QUEUEING PROCESSES WITH OPTIMIZATION BY ONE OR MORE DECISION-MAKERS

A THESIS

Presented to

The Faculty of the Division of Graduate Studies

By

Borge Tilt

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the College of Industrial Management

Georgia Institute of Technology

March, 1977

QUEUEING PROCESSES WITH OPTIMIZATION

BY ONE OR MORE DECISION-MAKERS

Approved:

K. R. Balachandran, Chairman

F. E. Williams

David Nachman

Robert B. Cooper

John A. White

Date approved by Chairman: MAR 2 1977

ACKNOWLEDGMENTS

Each member of the dissertation committee has made significant contributions to this work, for which I am grateful. The chairman, K. R. Balachandran, aroused my interest in optimization in queues and has given me much welcome advice and, not the least important, moral support when needed. I have profited from my many discussions with F. E. Williams who also assisted me in the design of a curriculum that has proven itself most helpful in my work. Robert B. Cooper has done a lot for my development into a queueing theorist, and coauthored Chapter V. David C. Nachman and John A. White have helped through their constructive criticism of the final draft.

Thanks are also due Claudine Taylor who did the typing, competently and cheerfully.

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

LIST OF TABLES

LIST OF ILLUSTRATIONS

SUMMARY

This dissertation consists of four independent studies. Its five chapters are: I. Introduction; II. Characterization of Queueing Processes with Optimization; III. Queues with Decision-Making by Individual Customers; IV. An M/M/1 Queue with Two Users Choosing Arrival Rates; and V. The Distribution of Maximal Queue Length in the M/G/1 Queue.

Chapter II develops a conceptual framework that systematizes the various elements of decision-making relative to queues. Treated in detail are: objective function and decision variables; decision points and action points; information structure; actions, decisions, strategies.

Chapter III discusses decision-making in queueing systems where individual customers have one or both of two options: (1) option to join or balk; (2) option to choose priority through a payment made on arrival. Each customer's objective is to maximize the expected net gain, defined as 0 if he balks, and as R-b-hE(W) if he joins, where R = reward, h = unit time waiting loss, b = payment, E(W) = expected waiting time. Considered are: (A) GI/M/s/N queue with either FIFO or LIFO rule, and balking option; (B) M/M/s/N queue with priority option; (C) M/M/s/N queue with balking and priority options. Customers limit their choice to a set of actions derived by considering the actions of other customers. If the set consists of just one action we call it optimal. A policy is defined as a decision rule selecting an action as a function of arrival state, R and h. The analysis focuses on optimal and stable policies. For model A we show an optimal policy exists, except, possibly, for $N=\infty$ with LIFO rule.

For model B we give conditions for a policy to be stable and induce either FIFO or LIFO service order. Similar results are derived in special cases of model C. An appendix discusses the effect of choice of dominance criterion in the process of reducing the action space, by deletion of dominated actions.

Chapter IV analyzes decision-making in an M/M/1 queue with two users, each choosing the rate λ_i (i=1,2) of a Poisson stream, attempting to maximize $R_i(\lambda_i)-h_i\lambda_i w(\lambda_1+\lambda_2;\mu)$, where $R_i(\lambda_i)$ = rewards per unit time, h_i = waiting loss factor, $w(\cdot)$ = mean waiting time in system at service rate μ. Following a treatment of the corresponding one-user model, the two-user model is analyzed under three different behavioral assumptions we can briefly characterize as (1) Mutual accommodation, (2) Domination, and (3) Cooperation. Our main purpose is to determine how arrival rates and profits depend upon behavioral assumptions.

Chapter V begins with a simple recursive formula for calculation of the mean busy period b_n of an M/G/1/n queue. Next, it is shown that in an M/G/1 queue that starts with i customers in system, the probability that at most k ($k \geq i$) customers will be present simultaneously during the i-busy period equals b_i/b_{k-i}.

CHAPTER I

INTRODUCTION

Queueing theory has its origin in attempts to solve real-world problems concerning waiting line phenomena. Yet, for many years, up until about ten years ago, nearly all papers in queueing theory dealt with queues without decision-making.

Recently, however, there has been an upsurge of interest in decision-making problems, especially those classified as optimal design and control of queues. Significant progress has been made in formulation of meaningful optimization models as well as in their analysis. Still it seems that little more than a beginning has been made in the analysis of decision-making in queueing systems. Many questions have been left unanswered, or have not been asked at all.

We propose to make two distinct contributions to the queueing literature: (1) A characterization of queueing processes with optimization (Chapter II); (2) An analysis of decision-making in certain queueing systems with more than one decision-maker (Chapters III and IV). The final chapter, Chapter V, is a byproduct of an attempt to generalize some results in Chapter III. All chapters can be read independently.

Chapter II discusses and classifies queueing processes with optimization. The classification is by defining characteristics, but only those which are peculiar to decision systems. Chapter II has been motivated in part by the apparent lack of a systematic exposition of the characteristics

of decision processes in queues. In addition to providing a useful frame of reference for queueing models, our analysis will also serve the purpose of giving direction to future research.

Models with multiple decision-makers form a large and largely unexplored area in queueing theory. Chapters III and IV are designed to partially fill the gap.

Chapter III treats, in much detail, some basic models in which individual customers are the decision-makers, with balking option or priority option. In the latter case, much space and effort are devoted to deriving conditions under which a policy, or decision-making rule, is stable and leads to either first-in-first-out or last-in-first-out processing of customers. Analyzed models are either new or generalized versions of models in the literature.

Chapter IV contains an analysis of an M/M/1 queueing system with two decision-makers, each choosing his arrival rate as a user. The model, which has not been analyzed before, is a special case of a duopsony model. Here, the arrival rate is equivalent to a quantity demanded per unit time. As a stepping stone to the analysis of the two-user model we analyze a model with a single user. In the two-user case we discuss the solutions under three different behavioral assumptions: (1) Mutual accommodation, (2) Domination, and (3) Cooperation.

Chapter V (published jointly with R. B. Cooper) deals with the distribution of the maximal queue length, k, during a busy period of an M/G/1 queue. It is found that the probability that at most k customers are present in case the queue starts with i customers equals b_i/b_{k-i}, where b_n is the mean busy period of an M/G/1/n queue (n = waiting room).

A similar formula is known in the literature. What is new, is the interpretation of b_i and b_{k-i}, as well as the simple derivation of the result.

The reading of the subsequent chapters requires a basic knowledge of standard queueing theory and its terminology, at the level of Cooper [1972] and Kleinrock [1975].

CHAPTER II

CHARACTERIZATION OF QUEUEING PROCESSES WITH OPTIMIZATION

Most of the queueing literature to date deals with queueing processes that do not explicitly recognize decision-making, and the sole purpose of the analysis is to derive performance characteristics for each queueing process as specified. In contrast, our concern is with processes which include decision-making by one or more decision-makers, and we focus on the decision process. Our objective is to develop a conceptual framework that systematizes the various elements of decision-making relative to queues.

Among other things, the here constructed framework should help in identifying meaningful and important decision models and should make it easier to relate existing, and as yet unformulated, decision models to one another. Much of what will be said applies to controlled stochastic processes in general. However, we need not, and shall not discuss the broader implications of the analysis.

In order not to overly complicate the exposition, it is confined to queues where each decision-maker has a single-valued objective function that must be maximized or minimized. Without loss of generality we assume maximization. In such cases we may speak of queueing processes with optimization, and optimization models. Also in the interest of simplicity we concentrate on queues with a single decision-maker. Only in the last section do we deal with queues with two or more decision-makers.

1. Concepts Current in the Literature

The question of characterization of optimization models has received little attention in the queueing literature. We shall mention only two recent survey papers which touch upon the subject. They are (1) Crabill, Gross and Magazine [1973], "A Survey of Research on Optimal Design and Control of Queues," and (2) Stidham and Prabhu [1974], "Optimal Control of Queueing Systems."

We choose to quote in full the appropriate sections of the two mentioned papers. The extracts present what the authors consider as important distinctions to make regarding optimization models. They also employ a widely accepted terminology. Hence they should serve well as a background for our discussion of the subject.

1.1. Quote from Crabill, Gross and Magazine [1973], pp. 1-2

> In the last 15 years there has been an increasing interest in the study of designing and controlling the behavior of queueing systems (prescriptive models as opposed to descriptive models which make up the majority of queueing literature to date). It is the purpose of this paper to describe briefly some of the important work performed in the prescriptive area. We shall concentrate on the models that have been considered and the results obtained rather than analytical techniques.
>
> Early work involving prescriptive models was concerned with design aspects of queueing systems. These were viewed as static optimization problems; that is, various system configurations were considered and the analysis of the resulting steady state behavior allowed one to determine the best system for optimizing some long run average criterion such as cost or profit. The term static model is used in the sense that once the system configuration is set, the system characteristics do not change over time. We will call models dynamic if the system operating characteristics are allowed to change over time. Most of the recent work in the optimization of queues has concentrated on dynamic models. These terms, static and dynamic, are intended only to provide a basis for a logical breakdown of some of the work discussed in this paper. Some models are a mixture of both static and dynamic considerations.

1.2. Quote from Stidham and Prabhu [1974]- p. 263

> Research in the control of queueing systems has been going on at an ever-increasing rate in the last few years. Progress in this field has reached the point where one can make some general statements about "the queueing control problem."
>
> Certain issues pervade the research done so far on queueing control. It is usually possible to categorize a paper by (i) the system structure (e.g., M/M/1, M/G/1, GI/M/s); (ii) the decision variables (e.g., service parameters (number of servers, service rate), arrival parameters (arrival rates, acceptance or rejection indicators for various classes), queue parameters (which class to serve next), operating-time parameters (when to close a queue)); (iii) the admissible decision epochs (e.g., arrival and/or departure epochs or all epochs: the choice usually being made so that the given system structure induces a Markov or semi-Markov decision process); (iv) the cost (e.g., prices for entering, rewards for being served, waiting costs, service costs, switching costs); (v) the objective function (e.g., discounted or average cost); (vi) the time horizon (finite or infinite).

1.3. Comments

The purpose of the two classification schemes has been a categorization of existing optimization models. Therefore it is not surprising that many interesting optimization models, whose analysis has scarcely begun, do not fit into the classification schemes, or that important special characteristics are unaccounted for. Examples are: queues with multiple decision-makers; with risk with respect to parameters of the system; with incomplete process information. Besides lacking in coverage and detail, the proposed classification schemes suffer from not being supported by a logical development of the elements of decision-making.

Our conceptual framework is based upon a study of the decision process rather than existing optimization models. The sole basis for our classification shall be the defining characteristics of the queueing process. Neither derived characteristics (i.e. the results of an analysis)

nor analytical method will have a part to play.

2. Concepts and Definitions

The present section begins by explaining the meaning of the words system and model in subsequent discussions. Next, we present a preliminary survey of the main elements of the decision process. We conclude with a discussion of risk and uncertainty.

2.1. System and Model

In the literature "system" often stands for something (perhaps just a "black box") which transforms input in a given way. In queueing theory it is convenient to define "system" differently. The convention that we shall follow is to let the input process be part of the system. Thus, for queues without decision-making the term system denotes a well-defined stochastic process. A simple example of a system in this sense is an M/M/1 queue (Poisson input, exponentially distributed service times, a single server) with arrival rate λ and service completion rate μ.

The term "model" is sometimes used synonymously with system as defined by us, but with the connotation that it is the approximate mathematical representation of a real-world phenomenon. However, when we speak of a model we mean a class of systems. An example of a model in this sense is the class of all M/M/1 queues such that $\lambda/\mu < 1$. Obviously, we will name as a model only a class of similar systems, that is, systems which differ only with respect to a number of characteristics which are or may be treated as parameters.

A study of a model may have as its sole objective the derivation of results concerning each system considered by itself. Frequently,

however, one will also be interested in a comparison of the systems, that is, measures of the effect of a change in system parameters.

We shall use the terms system and model also when certain aspects of a queueing process depend on a choice by a decision-maker. If so, we may speak of decision system and decision model.

In view of the implied similarity of the systems that comprise a model, we permit ourselves to speak of characteristics of a model, when we mean those of the systems.

2.2. Elements of the Decision Process

As far as the decision-maker is concerned the elements of the decision process are: decision points, decision spaces, information, strategy, and objective function. These concepts will now be briefly discussed.

2.2(a). Decision Points. The queueing process with its contained decision process evolves in time. The process has a beginning but not necessarily an end. At certain points in time the decision-maker must make a decision. These epochs are called decision points.

2.2(b). Decision Spaces. At each decision point a choice of action is required. The choice may be made directly, or it may be made indirectly, via a choice of a probability distribution over the available actions, in which case the action is drawn from the selected probability distribution. In general, we may assume that decision-maker chooses a probability distribution (possibly degenerate) and we call the selected distribution the decision. The set of all admissible probability distributions at the given decision point is termed the decision space. For technical reasons we may speak of choice, decision, and decision point,

even if only one action is available.

The two types of decision spaces completely dominating the literature are: (1) the set of all degenerate probability distributions, (2) the set of all possible probability distributions. Observe that a probabilistic choice of action may make sense even if there is only one decision-maker.

2.2(c). Information. The decision-maker does not act blindly. At each decision point he is given some information. This has two parts: (1) information about the queueing system itself (i.e. laws of evolution of the process), (2) information about the particular course of the process--the realization--up to and including the decision point (i.e. the history). Neither piece of information is necessarily complete. As far as process information is concerned it will be assumed that the decision space is always known. Throughout we make the assumption that all information received by decision-maker is true.

2.2(d). Strategy Space. By a strategy we shall understand a decision rule by which the decision at an arbitrary decision point is a function of the information available to the decision-maker at that point. By definition, under a given strategy all decision points with identical information will have identical decisions.

A strategy is admissible if it produces a decision at every decision point in every realization and if each decision necessarily lies in the proper decision space. An admissible strategy need not, of course, provide answers for situations (defined by the information) which cannot occur under the strategy. The set of all admissible strategies we call the strategy space and denote by Π. An admissible strategy is typically denoted π.

A decision-maker does not have to make a decision at a decision point until he arrives there. He need not make decisions ahead of time according to a strategy that takes into account every condition that later may arise. Yet, the strategy concept is indispensable in decision-making as, in principle, a decision-maker should make each decision with all possible futures in mind. There is no loss of generality in assuming, as we shall, that a strategy is selected already at the starting point of the process, say t_0.

2.2(e). Objective Function. The decision-maker's guide in making his decisions is the objective function, to be defined. The basic concept is the outcome, or realization. By that is meant the detailed description of the particular course taken by the queueing process. It is clear that the strategy choice π is an important determinant of the outcome. However, it is also affected by various random variables. Among these may be randomly selected actions. The process will always be affected by random variables which are inherent in a queueing process, namely interarrival times, service times and, perhaps, variables connected with availability of servers. In addition, other quantities which are typically constant, such as the arrival rate in a Poisson arrival process, may be random variables.

For a given $\pi \in \Pi$ the queueing process is a stochastic process. Suppose any particular outcome can be summarized into a single real number, the payoff, which completely describes the outcome as far as our decision-maker is concerned. The payoff, $w(\pi)$, generally is a random variable.

Let $F(\cdot;\pi)$ denote the distribution function of $w(\pi)$, which is allowed to be degenerate. Suppose that the distribution $F(\cdot;\pi)$ by the

decision-maker is summarized into a number, $v(\pi)$, by some formula, say $v(\pi) = f(F(\cdot;\pi))$, $\pi \in \Pi$. Our assumption is that $v(\pi)$ measures the worth of the choice π, and $v(\pi)$ is called the objective function value. A simple, but important example is that $v(\pi)$ is defined as the mean of $w(\pi)$. More generally, $v(\pi)$ may be defined as any function of the outcome distribution. The objective function is $v(\pi)$ considered as a function of the strategy π.

The decision-maker's objective shall be the maximization of $v(\pi)$. Normally, one would assume that the decision-maker knows all admissible strategies and that, at the starting point, he is given sufficient information about the system to enable him, at least in principle, to determine an optimal strategy. A detailed discussion of what constitutes sufficient information is beyond the scope of our analysis.

2.3. Risk and Uncertainty

Decision systems fall naturally into two classes: (1) systems with risk, (2) systems with uncertainty. We speak of a system with risk if, for each $\pi \in \Pi$, $w(\pi)$ follows a distribution which may be calculated, with any accuracy desired, from the information available to the decision-maker at t_0. We allow $w(\pi)$ to be constant for each π. On the other hand, a system with uncertainty is a system where, for some $\pi \in \Pi$, the distribution of $w(\pi)$ cannot be calculated by the decision-maker.

It has been assumed that all information about the queueing process is true. Hence, if we have a system with uncertainty, then the decision-maker must be uncertain about the system itself. That is, there is at least one defining characteristic whose distribution function is not completely known by the decision-maker, or some feature of the system is

incompletely described.

Knowledge of the distribution of a parameter is here taken to mean one of two things. Either the distribution is given and the decision-maker is informed thereof, or the parameter value is unknown, but the decision-maker has assumed a probability distribution and acts on that assumption. In the latter case the decision-maker has substituted a subjective, or a priori, distribution for the "unknown" quantity or feature.

Systems with uncertainty will not be considered in the following. We go a small step farther and make the assumption that the decision-maker has no uncertainty with respect to the defining characteristics.

3. Decision Points

Sections 3 to 7 treat in some detail the elements of decision-making which were reviewed in Section 2. In the present section our subject is decision points and certain closely related events.

3.1. The Set of Decision Points

Decision points may be of different kinds, differentiated by type of information (including decision space information) and, perhaps, the particular values describing the history. Here we are concerned with the distribution over time of one kind of points. Typically, it will suffice to distinguish one or two kinds.

Let $P(\pi)$ designate the set of decision points realized under strategy π. $P(\pi)$ may be a fixed set or a variable set. It is natural to distinguish four cases: (1) $P(\pi)$ is fixed for each $\pi \in \Pi$, and $P(\pi_1) = P(\pi_2)$ for $\pi_1 \in \Pi$, $\pi_2 \in \Pi$; (2) $P(\pi)$ is variable for each $\pi \in \Pi$, but $P(\pi_1)$ and $P(\pi_2)$ have identical probability distributions for $\pi_1 \in \Pi$, $\pi_2 \in \Pi$; (3) $P(\pi)$ is fixed for each $\pi \in \Pi$, but $P(\pi_1) \neq P(\pi_2)$ for some $\pi_1 \in \Pi$,

$\pi_2 \in \Pi$; (4) other. The queueing literature offers interesting examples of all four cases.

Case 1 is exemplified by systems with decision points as follows: (a) $P = \{t_0\}$ (a pure design system); (b) $P = \{t_0, t_0+\delta, t_0+2\delta, \ldots\}$ (a periodic review system); (c) $P = [t_0, \infty)$ (a continuous-time decision system). Case 2 is illustrated by a system with constant Poisson arrival rate and in which the decision points are all the arrival points. A simple example of case 3 is a system with decision points $P(\delta) = \{t_0, t_0+\delta, t_0+2\delta, \ldots\}$ where the review interval δ is to be decided at t_0. For an example of case 4 consider an M/M/1 system with variable service rate, whose level is decided at arrival and departure points on the basis of the current number of customers in the system.

3.2. Reaction and Implementation Times

A decision point may be predetermined by the definition of the system. It may also have been set at a previous decision point. Commonly, however, decision points are generated by "events" such as an arrival. In general, in these cases, time passes from the instant at which the event takes place until the decision is made at the decision point. This delay is called the reaction time. Sometimes the reaction time of a model can be interpreted as an information lag.

In nearly all queueing models the reaction time is assumed to be zero. That may be an oversimplification of real-world systems. Quite as reasonable, but also much harder to analyze, are models where the reaction time is assumed to be a positive constant or a random variable.

A decision is a deterministic or probabilistic choice of action. The usual assumption is that the indicated action is carried out

immediately. In general, the action will be carried out in one or more steps, or even continuously, following the decision. We shall consider only cases with a single action point associated with a decision point. The time between decision and action we call the implementation time.

Obviously, models with positive implementation times, be they constant or variable, are much more difficult to analyze than models with zero implementation times. In part this is due to the complication that a new decision point may occur before a previous decision has been implemented. Interestingly, in inventory theory models with separate decision and action points have been known for a long time whereas in queueing theory such models are rare.

4. Decision Spaces

Our discussion of decision points dealt with the question of when decisions are made. Now we turn to the question of what shall be decided at a decision point according to the system specification. For simplicity we limit ourselves to a deterministic choice of action. A generalization to probabilistic choice can be accomplished without difficulty. Also, for simplicity, implementation time zero is assumed.

4.1. The Decision Space

In the case of a deterministic choice, the decision space is, in effect, the same as an action space, namely, a set of (admissible) actions, from which the decision-maker can choose freely.

Usually the action space is directly specified. An indirect specification is also possible. Then actions are termed admissible and may be selected if they belong to some given set and meet certain conditions.

An example is the choice of number of servers under the restriction that future customers with probability P or greater wait at most T units of time for service to begin.

An action may concern any aspect of a queueing process which can reasonably be assumed to be under the control of the decision-maker. In the queueing literature most actions deal with physical characteristics of the service facility or customer flow characteristics. There are many ways actions can be categorized. We shall not, however, discuss that subject.

Denote by A the set of admissible actions (or alternatives) at the decision point and let $\underline{a}$ denote an arbitrary action in A. The action may be multidimensional. That is, it concerns $n \geq 2$ quantities whose levels are set simultaneously. If so we may write $a = (a_1, a_2, \ldots, a_n)$. An important special case is that the n characteristics can be chosen independently. Then A can be written $A = A_1 \times A_2 \ldots \times A_n$, with $a_j \in A_j$, $j = 1,2,\ldots,n$.

If the action space A is one-dimensional and quantifiable, then the action, viewed as a variable, is usually called a decision variable. It is natural to extend the use of this term to each component of $a \in A$ in case $A = \times A_j$.

4.2. The Collection of Decision Spaces

Until now only the decision space at a single, arbitrary decision point has been the subject. A characterization of the whole collection of decision spaces of a realization is also of some interest. Our discussion of this subject is limited to the concept of a stationary decision space.

A system is said to possess a stationary decision space if all

decision points, except possibly t_0, have the same decision space. Similarly, we say a system has a class-dependent, stationary decision space if the decision points fall into classes within which all points have the same decision space, and, furthermore, the date is not used in the classification. We are interested only in cases where the classes are defined by information given to the decision-maker, excepting the decision space itself.

5. Information

Although information plays a crucial role in decision-making, very little is said in the literature about information. Perhaps the reason is that most models implicitly assume an omniscient decision-maker.

To us, only information available at decision points is relevant. Generally, the information at a decision point has two components, namely system information and process information. The two kinds of information will be discussed in turn.

5.1. System Information

The system has been defined as the complete set of laws governing the evolution of the queueing process under consideration. The description of the system must be all encompassing. Among the things specified are the mechanism for generation of decision points as well as the information available at each individual decision point. System information is simply information about the system as specified.

An important special case is when the system information is the same at all decision points. In that case we may as well suppose that the information is available from the beginning, t_0. In the following we assume constant system information.

5.2. Process Information

Consider a realization of the queueing process. Denote by H(t) the description in complete detail of the outcome until time t, $t \geq t_0$. If t is a decision point then H(t) will include the current decision space, but not the decision. Assuming that t is a decision point, let I(t) denote the information available to the decision-maker at t about H(t). Thus I(t) is a subset of H(t) or has been derived from H(t). We require that I(t) always states the current decision space. I(t) will also be referred to as the process information at t.

The subsequent discussion of process information concerns (a) Choice of information; (b) Contents; (c) I(t)'s relation to H(t); (d) Relation between I(t)'s.

5.2(a). Choice of Information. The system may leave it to the decision-maker to decide at some decision point what information should be made available to him, at present or future decision points. For instance, he might have to choose between number of customers in the system and total service time required of all customers present. Typically, different costs will be associated with different choices of information basis as well as the collection of that information. If decision-maker exercises control of the information given him we shall speak of variable information, and otherwise fixed information.

5.2(b). Contents. Process information may deal with virtually any aspect of the process. A categorization of all items of information will not be attempted. Two categories that suggest themselves are facility information and customer flow information. In a special category one finds information relating to past decision points, concerning the information

available (data, or kind only), decisions made, and action taken.

5.2 (c). I(t)'s Relation to H(t). I(t) may range all the way from the minimal, where only the decision space is given, to the maximal, H(t), where decision-maker is told everything that has happened. Note that while few queueing models prescribe maximal information, frequently the information I(t), though less than H(t), is as good as H(t) for the purpose of maximizing the objective function.

In the intermediate cases, I(t) is more than just decision space at t, but less than H(t). Typically, I(t) is the result of both a reduction and a summarization of the data contained in H(t). Reduction may take the form of unavailability of data in H(t), caused by time lags in the information process.

Frequently I(t) less current decision space consists of a vector of data of the same type for all decision points, except possibly t_0. In such cases the vector is commonly called the state of the system. The data may represent current characteristics of the process at t, as number of customers in the system, total residual service time (work load), whether a particular server is available or not. However, some data may also be based on the past history. Examples are total number of arrivals or operating costs to date. The set of all possible vectors is often called the information space.

5.2(d). Relation between I(t)'s. An interesting and significant aspect of the information structure is the relation between I(t)'s at different decision points of a realization. We are especially interested in the degree of loss of information about the process as time passes. The question will be discussed only under the assumption that the decision-

maker is informed of either the exact date or the decision point number.

We begin by defining systems with no loss of information. Let t_1 and t_2 ($t_1<t_2$) denote two decision points in a realization of the process. We shall say there is no loss of information between t_1 and t_2 if and only if the decision-maker at t_2 knows $I(t_1)$ exactly. Observe, it is not sufficient that $I(t_1) \subset I(t_2)$, or in other words that everything known at t_1 is also known at t_2. This weaker condition means only that the process information is refined and extended in time. We say that the system has no loss of information provided there is no loss of information between any two decision points, whatever decisions are made.

At the other extreme one finds the systems with a total loss of information. Here nothing is known at a decision point about the process information given at previous decision points, not even the date, except what might be deduced from the system information.

Between the two extremes are a variety of systems with, at least potentially, a partial loss of information.

5.3. Design versus Control Systems

The information structure plays an important part in our distinction between design systems and control systems. Notice that our concept of design and control systems differs somewhat from others appearing in the literature.

A design system is defined as a decision system with the following two properties: (1) Decision point dates (as measured by the exact time or the sequence number) are fixed, and the decision space is a function of decision point date; (2) The process information consists of only decision point date and associated decision space. Thus, in a design system

the decision-maker will never be told anything not known at t_0. As a consequence, optimal decisions can be made, unconditionally, already at t_0, for all decision points.

An obvious example of a design system is a system in which the only decision is made at t_0, concerning, say, the number of servers for all time. For another example, consider a continuous-time decision system where the arrival rate is a known function of time, the decision variable is the number of servers, and no process information except date and common decision space is available.

By a control system is meant a decision system which is not a design system. Here, typically, some process information will be available which precludes the determination of a sequence of optimal decisions already at the start of the process.

A comparison of our design and control system concepts with others in the literature is interesting. Unfortunately, these and the related terms, static and dynamic systems, or models, are often ill-defined or not defined at all. Little will therefore be said on the subject. We limit ourselves to a comment on the quoted paper by Crabill, Gross and Magazine [1973]. They make a distinction between static models, with constant "system operating characteristics," and dynamic models. To us, this distinction does not seem so important. What is needed in the case of decision models is a distinction which, like ours, is based on the information structure.

6. Strategy Space

Recall that the strategy space is the set of available strategies, and that a strategy gives a decision as a function of the information at

a decision point. One obvious descriptor of a strategy space is the number of strategies, in particular whether finite or infinite. Further characterization rests on a characterization of individual strategies. We have mentioned the classification of strategies into randomized and nonrandomized strategies. Here, our subjects are the class of stationary strategies and the subclass of static strategies. Often, in the literature, all strategies in the strategy space belong to one of these classes.

6.1. Stationary Strategies

We call a strategy stationary if the decisions do not depend on information on the date of the decision point, except possibly at t_0, whether such information is available or not. A simple illustration is a strategy making the decision a function of number of customers in system and current number of activated servers.

6.2. Static Strategies

An interesting group of stationary strategies are formed by strategies whereby the decision is a function of decision space only. Among these strategies are the static strategies, to be defined. The term may not be too fortunate, but it has gained some currency in the literature. An underlying assumption is that there exists a strict, simple ordering of all possible decisions, or actions, for each decision point. Thus, all decisions in a decision space are ranked so that (1) of any two decisions, one is ranked the higher, and (2) if decision 1 is ranked higher than decision 2, which in turn is ranked higher than decision 3, then decision 1 will be ranked higher than decision 3.

A static strategy is a strategy which at every decision point, except possibly t_0, selects the one available decision which is ranked the

highest. Consider for example queueing systems where the decision-maker may select for service any customer in the queue. Clearly, FIFO and LIFO service orders qualify as static strategies, with customers ranked by arrival time or number. Suppose the customers belong to predetermined priority classes 1,2,... . Then the strategy (as in the simple priority queue) of selecting the customer who is in the highest priority class represented and arrived the first, is static.

7. Objective Function

Our subject is not the objective function $v(\pi)$ as such, that is, $v(\pi)$'s dependence on π. Rather, our concern is the expression for $v(\pi)$ or for the payoff on whose distribution $v(\pi)$ is based.

Typically, the payoff is given as a simple expression in (a) system constants, (b) decision variables, and (c) performance variables. We are interested in the form and the interpretation of this expression. In Sections 7.1 and 7.2 we discuss system classifications based upon properties of the expression for the payoff, and hence characterize the objective function expression. In Section 7.3 we discuss economic systems as opposed to physical systems, with emphasis on the objective function. Finally, in Section 7.4 we discuss a categorization of decision-makers, based solely on an interpretation of the expression for the objective function value.

7.1. Planning Horizon

Consider a realization under strategy π. Suppose T is a point in time such that the exact payoff can be determined at T, but not before. T is called the planning horizon. In general, T is a random variable

which is dependent on π. Possibly, $T = \infty$.

Systems with infinite horizon ($T \equiv \infty$) are predominant in the queueing literature and need not be discussed. Among other possibilities we shall consider only systems with the property that for no $\pi \in \Pi$ can $T = \infty$ occur. Such systems we call finite horizon systems. The following distinction of four categories is natural: (1) T is independent of π and constant (fixed horizon); (2) T is independent of π, but is a random variable; (3) T is dependent on π and constant for each $\pi \in \Pi$; (4) T is dependent on π and, at least for some π, a random variable. In each category, interesting examples are easily found.

7.2. Additivity Properties

We shall discuss two additivity properties often encountered in optimization models, either singly or in combination.

First we define additivity with respect to performance variables. Let n be the number of performance variables. For given $\pi \in \Pi$ let $x_0(\pi)$ be a constant and let $x_1(\pi), x_2(\pi), \ldots, x_n(\pi)$ denote the n performance variables, which may themselves be expressions involving more basic performance variables. We say the payoff is additive in the performance variables if

$$w(\pi) = x_0(\pi) + \sum_{i=1}^{n} x_i(\pi).$$

This property is particularly useful in case the objective function value equals expected payoff.

Second we define additivity with respect to time. Let $\pi \in \Pi$ and let $t_0, t_1, \ldots, t_m$ ($m \leq \infty$) be a partition of the time scale, which may be

randomly determined or, perhaps, even arbitrary. Let $y_0(\pi)$ be a constant, and denote by $y(t_{j-1},t_j;\pi)$ a partial payoff associated with interval no. j, and which depends on the history of the process before t_j. We say the payoff is time additive if

$$w(\pi) = y_0(\pi) + \sum_{j=1}^{m} y(t_{j-1},t_j;\pi).$$

It is easy to see that a system possesses both of the mentioned additivity properties if we have

$$w(\pi) = k(\pi) + \sum_{i=1}^{n} a_i(\pi) + \sum_{j=1}^{m} b(t_{j-1},t_j;\pi)$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{m} z_i(t_{j-1},t_j;\pi),$$

where $k(\pi)$ is an overall constant; $a_i(\pi)$ is a constant associated with performance variable no. i (i=1,n); $b(t_{j-1},t_j;\pi)$ is a constant associated with interval no. j (j=1,m); $z_i(t_{j-1},t_j;\pi)$ is the contribution of interval no. j to performance variable no. i.

7.3. Economic Interpretations

Both queueing systems with and without decision-making may be divided into physical systems and economic systems. Typically, a system without a decision-maker is a physical system and has no objective function, while a system with a decision-maker must have an objective function and typically is an economic system. We will discuss only systems with a decision-maker.

By a physical system is meant a system whose defining characteristics are all physical variables, dealing for instance with number of units and time intervals, and their probability distributions. The objective function value is also required to be a physical variable.

Conversely, an economic system is a system in which at least one defining characteristic is an economic variable by interpretation. Examples of economic variables are variables with an interpretation as revenue, cost, profit, discount factor, or utility. We note that often the objective function is additive in performance variables interpreted as revenue or cost.

Physical decision systems may seem odd, but they do occur in the literature. However, they are often, implicitly, economic systems. Consider for example an M/M/1 system, where customers may be selected from the queue in any order, each customer's service time is known in advance, and the objective is to minimize the mean queueing time (a reformulation as a maximization problem is easy). The solution is to always select the customer with the shortest processing time (service time). No economic variables enter explicitly into the system definition. Yet the system is implicitly economic. If a fixed cost per unit time spent in queue by a customer were part of the system definition and the objective function were defined as this cost constant times the mean queueing time, then the optimal strategy would not change.

7.4. Types of Decision-Maker

A decision-maker is an actual or imagined individual, or collection of individuals. In the real world as well as in the literature the decision-maker commonly belongs to one of three types to be described.

The sole basis for the categorization is the definition of the objective function. Since it is of some interest what strategies are available, this question will also be discussed, for each type of decision-maker, but only in terms of decision variables.

7.4 (a). Facility Owner. The term facility owner is applied to an individual, a firm, or the like, that bears the cost of facilities and their operation, and in return derives revenue from the service of customers. Both cost terms and revenue terms enter the objective function and the payoff expression, which is increasing in revenue terms and decreasing in cost terms.

When the payoff equals the difference between revenue and cost, it is frequently named profit, and the facility owner is labeled profit maximizer. If cost terms are absent or independent of π, then the facility owner may be called a revenue maximizer. Similarly, if revenue terms are absent or independent of π, then he may be called a cost minimizer.

There is hardly any aspect of a queueing process which cannot, in one context or another, reasonably be assumed to be under the partial or complete control of the facility owner. Among the most important decision variables are, of course, the physical characteristics which define the facility, such as number of servers, waiting room and service rate. Another family of decision variables that may well be set by the facility owner are variables influencing the customer flow, directly or indirectly. Examples abound in the literature, but will not be given here.

7.4(b). Customer Agent. This term denotes a decision-maker whose objective function summarizes all benefits and costs of individual customers, among which are: reward from service, payment for service, waiting

loss. We shall not discuss how the collective welfare might be or should be measured. Suffice it to say that often the objective function is defined simply as the expected total or average payoff to individual customers, each of whom is endowed with his own payoff function.

The identity of the customers of interest to the customer agent is not always obvious. Certainly all arrivals should be included, and if the customer agent can influence in any way the stream of arrivals, then the class of customers to be considered should include potential arrivals as well. In the case of queues with sources for repeated requests for service--call them users--one might define user payoffs and calculate the agent's payoff from the user payoffs.

Perhaps the most important class of decisions under the purview of a customer agent are customer flow controls. These fall into two broad categories. In the first category fall decisions that may be thought of as having been delegated by the customers to their agent. Examples are balking and reneging options. In the second category fall other means of control. A good example is the agent's use of a toll, the decision variable being its level, with the purpose of discouraging entry into the system of certain customers who would otherwise cause congestion while profiting little themselves by joining rather than balking. Another example is an agent's freedom to decide the order of service.

Observe that the presence of a customer agent does not preclude the possibility that some decision, such as a balking option, is left with the individual customers who act to maximize their individual objective functions.

7.4 (c). Total System Optimizer. This term suggests a decision-

maker with an objective function which takes into account the interests of both facility owner and individual customers as described above. The decision variables under the control of the total system optimizer may be virtually any feature of a queueing system, and there is no need for their enumeration in this place. The difficult question is how the interests of the two parties are to be reflected in the objective function. We shall not examine this question.

8. Two or More Decision-Makers

As before, our interest is the characterization of queueing systems, not results, nor analytical methods. Evidently, the class of decision systems with two or more decision-makers is substantially more difficult to characterize than the class of decision systems with a single decision-maker. For this reason we shall make no attempt at a classification with respect to all important characteristics of this family of decision systems. Only a few observations on the subject will be offered.

A classification by number and type of decision-makers is useful. In this respect we distinguish three classes of decision systems. In class 1 are all systems with a fixed and finite number of decision-makers, such as a system with a facility owner as well as a customer agent, or a system with two users making repeated requests for service. In class 2 are the systems in which the set of decision-makers, though not fixed and finite, is independent of strategies followed. An example thereof is an M/M/1 system where the decision-makers are the facility owner (deciding service rate) plus every arriving customer (deciding whether to balk). In class 3 are the systems in which strategy choices will affect the

composition of the set of decision-makers. Either a decision-maker can directly create other decision-makers, or he can by his actions influence the mechanism for generation of decision-makers. An example of the latter is a system where a facility owner decides the size of the waiting room, and where the other decision-makers are those arrivals which can be admitted into the system.

A full description of a multiple decision-maker queue will contain largely the same elements as does the single decision-maker queue. In particular, the decision process for each decision-maker is formally little changed by the existence of other decision-makers. One difference is that each decision-maker is given some information concerning the other decision-makers. It goes without saying that the presence of other decision-makers may vastly complicate the decision-making for all.

A further characterization of these decision-systems shall not be attempted however important and interesting the subject.

In conclusion we shall offer a few remarks on the possibility of a solution, that is, a determination of an optimal strategy by each decision-maker, a most important issue.

Typically, in systems with multiple decision-makers no one can determine an optimal strategy. The reason is, of course, that the very existence of a strategy choice by other decision-makers typically constitutes an element of uncertainty which precludes an exact calculation of objective function value for some or all $\pi \in \Pi$. There are, however, exceptions to the rule.

Consider for instance a decision-maker, D, with complete system information. D will be in a position to select an optimal strategy under

a variety of conditions among which are: (1) The strategy choices by other decision-makers are not relevant; an example is an M/M/1 system with FIFO service rule, where D is an arbitrary customer, a facility owner decides arrival rate at t_0, and each customer decides whether to join or balk, informed only of number of customers present on arrival; (2) D will be informed of other decision-makers' strategy choices before he is required to select his strategy; (3) D will be informed of decisions or actions by other decision-makers and this knowledge suffices for identification of an optimal strategy; (4) D can deduce the (optimal) strategies by all those decision-makers who may affect his objective function value.

CHAPTER III

QUEUES WITH DECISION-MAKING BY INDIVIDUAL CUSTOMERS

Queueing models letting individual customers decide whether to join the system or balk, or letting them choose their own priority have appeared in the literature only recently. Among the papers dealing with aspects of customer decisions are: Adiri and Yechiali [1974], Balachandran [1972], Balachandran and Lukens [1976], Balachandran and Schaefer [1975], Kleinrock [1967], Knudsen [1972], Naor [1969], and Yechiali [1972].

The present chapter is most closely related to Balachandran [1972]. Like his work, this chapter focuses on stable policies that result in a given service order. Here we include the balking option besides the priority option, and we deal with FIFO queues in addition to LIFO queues. Moreover, the objective function parameters are allowed to be random variables, and the queueing system is generalized from M/M/1 to M/M/s/N and in some cases to GI/M/s/N.

Our analysis concerns individual decision-making in three basic models to be described in detail in Section 1. Model A is a GI/M/s/N queue with balking option. Model B is an M/M/s/N queue with priority option. Model C is an M/M/s/N queue with balking option and priority option.

Our aim is twofold: (1) to put the analysis of individual decision-making in our and related models on a firm theoretical basis; and (2) to derive interesting and useful results for the important basic models we

call A, B, C.

The underlying difficulty in the analysis of Models A, B, C is that, generally, one cannot derive customer behavior from the defining characteristics of the model. (The important exceptions are for Model A.) For this reason the concept of behavioral assumptions and the associated concept of a stable policy have been introduced into the analysis. The inclusion of behavioral assumptions on the part of the customers removes a residual uncertainty in decision-making due to the interaction of customer decisions.

In Section 2 we develop the theoretical framework for dealing with decision-making by individual customers. The key concepts are: complete action sets, reduced action sets, optimal policy, stable (i.e. self-sustaining) policy.

In Sections 3, 4, and 5 the mentioned concepts are used in exploring decision-making in Models A, B, C, respectively. Model A, for the cases of FIFO rule, and LIFO rule with $N < \infty$, always has an optimal policy. In all other cases, an optimal policy does not generally exist, and the questions of existence and uniqueness of stable policies still await their complete resolution. However, for Models B and C we do have some interesting results concerning conditions under which a policy is stable and at the same time induces FIFO or LIFO service order.

1. Model Specification

We consider a system with $s \geq 1$ servers and a single waiting line (queue), with room for N customers in the queue. Service times follow the exponential distribution function $H(t) = 1-e^{-\mu t}(t \geq 0)$. Customers arrive at intervals that are independent, identically distributed random

variables following either a general distribution function G(t) ($t \geq 0$, $G(0) = 0$) with finite mean λ^{-1} (Model A), or the exponential distribution function $F(t) = 1-e^{-\lambda t}$ ($t \geq 0$) (Models B and C). Customers arriving when the waiting room is full are turned away, never to return. Once accepted into the system a customer stays until served. Service is nonpreemptive. In order to avoid certain trivial complications we assume that the system starts empty.

A customer receives a reward $R > 0$ if served, but must make some nonnegative payment when joining the system, and also suffers a waiting loss equal to $h > 0$ per unit time spent in the system. If a customer does not join the system, and consequently is not served, he receives a reward $r < R$. Without loss of generality we set $r = 0$. The pair (R,h) is a random variable on the population of customers, say with distribution function U(R,h). The set of possible values of (R,h), the sample space, is denoted by Ω, and we let H denote the set of possible values of h. Thus $H = \{h:(h,R) \in \Omega \text{ for some } R > 0\}$. A customer's net gain, V, shall be defined as follows: $V = r \equiv 0$ if the customer does not join the system; $V = R-b-hW$ if the customer joins, with b being his payment, however determined, and W being his actual (total) waiting time, i.e. the time from arrival until completion of service. Each customer's objective is maximization of his expected net gain, E(V).

Usually in queueing theory actions by the customers are prescribed in every detail by the model. Here we explicitly allow individual customers some measure of influence upon their expected net gain through a decision to be made at the time of arrival. In Model A we give the customers a balking option, in Model B a priority option and in Model C both

of these options. A balking option is the right of an arriving customer to join or balk (not join) as he sees fit. A priority option is the right of a customer who joins the queue to choose any priority payment b from a given set B of nonnegative payments, with the understanding that priority payments determine the order of service of those waiting in the queue; high payment gives priority over low payment, and FIFO (first-in, first-out) rule applies to customers who have made the same priority payment. First-out is interpreted as first out of queue (into service).

Since Model A has no priority option, the order of service of customers in the queue must be specified: We shall assume either FIFO or LIFO (last-in, first-out) rule. Likewise, since Model B has no balking option the conditions under which a customer joins the system need to be specified: All arriving customers will be assumed to join, waiting room permitting. This rule we refer to as "forced joining".

We are interested mainly in situations in which customers act under assumptions that permit calculation of E(V) for each available action. It may happen that two different actions result in the same value of E(V). Such ties are resolved by the following preference rule: A customer with a balking option prefers to join rather than balk, if the expected net gain by joining equals r = 0, and a customer with a priority option prefers the smaller payment if two priority payments yield the same expected net gain. Thus, outcomes are ordered lexicographically, with E(V) as the primary criterion and the action itself as the secondary criterion.

Further model specification necessitates some additional notation. Let the variable n denote the number of customers present in the system just prior to the arrival of a customer. Clearly, $0 \leq n \leq s+N$. For

notational convenience we define $\nu = n-s+1$ and refer to ν as the arrival state. Thus $-s+1 \leq \nu \leq N+1$. (Note that an arrival state $\nu \geq 1$ indicates $\nu-1$ customers in queue with all servers busy.) The term (ν,R,h)-customer shall mean a customer who arrives in state ν with objective function parameters R and h. Similarly, let a ν-customer be any customer arriving in state ν.

Each customer must make some nonnegative payment b upon joining the system. If a priority option is not available (Model A), or $\nu \leq 0$, then we will assume that the payment for joining is a function of ν, R and h, say $b(\nu,R,h) \geq 0$. If, on the other hand, a priority option is available and $\nu \geq 1$, then the customer can freely choose his payment from the set B. However, in the latter case we may be studying a policy whereby customers, voluntarily, pay an amount that is a function of ν, R and h, and then we will extend the function notation $b(\nu,R,h)$ to denote a policy payment as well.

The information that our models make available to an arriving customer has two distinct components: private data and system information.

Private data are variables that relate specifically to the customer himself, namely the number of customers observed in the system at arrival, measured by ν, and his objective function, defined by R and h. In what follows we denote the private data by the triple (ν,R,h).

The system information, on the other hand, is public data which is shared by all. It consists of the complete knowledge of the queueing system as described, including parameter values, distribution functions and option(s), plus, of course, the information that every customer is given the same kind of information he has, both the private data and the

system information. Thus, the words "system information" are meant to summarize the general knowledge about the queueing system, except decision-making.

2. Customer Behavior

The analysis of customer decision-making naturally falls into two parts. First we deduce as much as we possibly can about the decision-making. That is, we assume that the information upon which customers base their decisions consists of private data and system information only, and we ask what actions various categories of customers will or will not take. If an optimal action can be determined for each arriving customer, then the decision problem has been solved. However, there are cases where not every arrival has an optimal action. For further analysis of such cases we introduce behavioral assumptions on the part of the customers.

The key concepts of our analysis are: complete action sets, reduced action sets, optimal policy, and stable policy. These and related concepts are perhaps easier to understand if first we discuss the corresponding concepts of a simpler model than ours, namely the n-person game with pure strategies.

2.1. An Analogy

Consider a noncooperative n-person game with arbitrary sets of pure strategies to be termed actions. Let $v_k(a_1,\ldots,a_n)$ designate the payoff to player k when player ℓ selects action a_ℓ from the set A_ℓ^1, $\ell=1,\ldots,n$. Each player knows all n payoff functions. No player has prior knowledge of actions taken by others. Mixed strategies are not admissible. What will the players do?

We suppose each player attempts to maximize his payoff. This

objective is interpreted as follows. Let player k assume that the other n-1 players will select an action combination $t \in T$, where T is some set of combinations of available actions. No $t \in T$ is ruled out. Denote the conditional payoff function by $v_k(a_k;t)$. Let there also be given a preference rule which, for each player, provides a strict, simple ordering of actions, as in models A, B, C. That is, all actions are ranked so that for any two actions one is ranked higher than the other, and if a_k' is ranked higher than a_k'', while a_k'' is ranked higher than a_k''', then a_k' will be ranked higher than a_k'''. Given a choice between $a_k' \in A_k^1$ and $a_k'' \in A_k^1$ we shall assume the player will choose a_k', and we say a_k' <u>dominates</u> a_k'' if $v_k(a_k';t) \geq v_k(a_k'';t)$ for all $t \in T$ and, in case $v_k(a_k';t) = v_k(a_k'';t)$ for any $t \in T$ then a_k' is preferred to a_k''.

Note that our dominance concept differs from the usual dominance concept by which a_k' dominates a_k'' if $v_k(a_k';t) \geq v_k(a_k'';t)$ for all $t \in T$ and $v_k(a_k';t) > v_k(a_k'';t)$ for some $t \in T$. When we need to distinguish the two, we call ours strong dominance and the usual concept is termed normal dominance. These concepts as well as a third concept, strict dominance, requiring $v_k(a_k';t) > v_k(a_k'';t)$ for all $t \in T$, are discussed in Section 6. Notice, strong dominance, as opposed to normal dominance, implies preservation of dominance under reduction of T. This fact works in favor of the strong dominance concept.

Player k's <u>complete action set</u>, A_k^1, is the set of actions available to him. Frequently some of these actions can be ruled out by a successive dominance argument.

One version of this argument, simultaneous reduction--discussed by Luce and Raiffa [1957, p. 108] for the case of mixed strategies--runs as

follows. First, considering $v_k(a_1,\ldots,a_k,\ldots,a_n)$ defined for $a_\ell \in A_\ell^1$ (ℓ=1,...,n), player k (k=1,...,n) eliminates all dominated actions. The result is the partially reduced action sets $\{A_\ell^2\}$. Now, each player need consider only $\{A_\ell^2\}$. An examination of $v_k(a_1,\ldots,a_k,\ldots,a_n)$ defined on $a_\ell \in A_\ell^2$ (ℓ=1,...,n) may reveal new dominated actions, resulting in the second-stage reduced action sets $\{A_\ell^3\}$. Continuing, one derives for each player k a decreasing sequence $A_k^1, A_k^2, \ldots$. Let $A_k^* = \lim_i A_k^i$. We call A_k^* the reduced action set. Conceivably, if A_k^1 is an infinite set, A_k^* may be empty.

Another natural reduction scheme is cyclical reduction, by which the players in cyclical order, say 1,2,...,n,1,2,... reduce their action sets as much as possible.

Many different reduction schemes may be devised, including some according to which some players are never called upon to reduce their action set. Clearly then, $\{A_k^*\}$ may depend on the choice of reduction scheme. However, some reduction schemes are of little interest to us and will not be considered. Let an admissible reduction scheme be defined by two requirements: (1) there is a finite upper limit on the number of consecutive steps at which a player does not reexamine his actions, and (2) on his turn a player eliminates all dominated actions.

In Section 6 we show that under our dominance criterion, strong dominance, $\{A_k^*\}$ is unique, that is, the same for all admissible reduction schemes, for arbitrary $\{A_k^1\}$.

If A_k^* consists of a single action a_k^* then we term it player k's optimal action. If each player has an optimal action, then we shall say that the action combination $(a_1^*,\ldots,a_n^*)$ is an optimal solution.

Define a stable solution (equilibrium point) as any combination of actions $(a_1^o,\ldots,a_n^o)$ such that, for each k, a_k^o dominates all $a_k \in A_k^1 - a_k^o$ on $t_k^o = (a_1^o,\ldots,a_{k-1}^o,a_{k+1}^o,\ldots,a_n^o)$. Thus, by our definition of dominance, $(a_1^o,\ldots,a_n^o)$ is (strongly) stable if $v_k(a_1^o,\ldots,a_k^o,\ldots,a_n^o) > v_k(a_1^o,\ldots,a_k,\ldots,a_n^o)$, or $v_k(a_1^o,\ldots,a_k^o,\ldots,a_n^o) = v_k(a_1^o,\ldots,a_k,\ldots,a_n^o)$ but a_k^o is preferred to a_k, for all $a_k \in A_k^1 - a_k^o$, all k.

It is shown in Section 6 that with our definitions every stable solution (a_k^o) lies in the reduced action space, that is, $a_k^o \in A_k^*$, all k. We also show that an optimal solution is stable. Hence an optimal solution is the uniquely stable solution.

In Figure III-1 we illustrate with two-person games with three actions for each player. These games have been designed to highlight the distinction between optimal and stable solution. The preference rule plays no role here.

(i) Optimal, one-step solution

2	3	4
1	2	3
0	1	2

(ii) Optimal, multi-step solution

2	3	3
1	4	0
1	2	4

(iii) Stable, not optimal solution

2	3	3
1	4	0
1	0	4

Figure III-1. Solution Concepts in Games with Pure Strategies

As usual, the entry in row i and column j is the payoff to player 1 (loss for player 2) when player 1 chooses his action no. i and player 2 chooses his action no. j. Each game has a single stable solution (i,j) = (1,1) with payoffs (2,-2). In games (i) and (ii) the solution is optimal.

Assuming simultaneous reduction, the optimal solution in game (i) is identified in one step whereas in game (ii) four steps are needed to fully reduce the action sets and identify the stable solution as optimal. In game (iii) a stable solution exists but no reduction of the action sets is possible.

In our application the distinction between optimal and stable solutions is of fundamental importance. The main difference between the two is that an optimal solution follows from the model itself, whereas a merely stable solution will be effected by the decision-makers only under behavioral assumptions that go beyond the model specification.

2.2. Decision-Making in Models A, B, and C

The conceptual apparatus developed for the analysis of decision-making in games with pure strategies applies to Models A, B, and C, with minor modification, due mainly to the fact that our models are not characterized by a fixed and finite number of players as is the n-person game. Now, rather than n players we deal with a number of categories of potential customers, each category being defined as customers with the same (ν,R,h).

In all our decision models the complete action set for a (ν,R,h)-customer is a function of ν only, say $A(\nu,R,h) \equiv A_\nu$. Let $A = \cup A_\nu$.

The reduced action sets are derived by the inductive reasoning process described in Section 2.1. Denote the unique reduced action sets by $\{A^*(\nu,R,h)\}$. Observe that $A^*(\nu,R,h)$ is the same for all (ν,R,h)-customers. In Section 3.2(c) we shall present an actual derivation of $\{A^*(\nu,R,h)\}$ for for a special case of Model A, which exhibits a fair degree of complexity. A similar approach can be employed in other models. In case $A^*(\nu,R,h)$ has a single element we denote the action by $a^*(\nu,R,h)$, and call it the

optimal action for (ν,R,h)-customers.

Surely, a (ν,R,h)-customer chooses his action from the set $A^*(\nu,R,h)$ (unless $A^*(\nu,R,h) = \emptyset$). On the other hand, the model specification does not permit further reduction.

At this point we pause to define a policy and its balking limit:

DEFINITION (Policy). A policy is a stationary, nonrandomized decision rule which is used by all arriving customers and selects an action depending only on (ν,R,h).

The words "all arriving" imply there is a balking limit ν_o, $-s+1 \leq \nu_o \leq N$, such that an action is specified for all ν-customers with $\nu \leq \nu_o + 1$, where for each $\nu \leq \nu_o$ the prescribed action for at least some $(R,h) \in \Omega$ is to join, and for $\nu = \nu_o + 1$ the prescribed action, for all arrivals, is not to join.

By its definition, a policy need not specify an action for $\nu > \nu_o + 1$, assuming, of course, that every customer follows the policy. A complete policy is a policy that specifies an action for all (ν,R,h), that is, for all $\nu \leq N+1$, all $(R,h) \in \Omega$.

If all arriving customers have an optimal action, then the decision problem is automatically solved. Again, there exists a balking limit ν_o and $a^*(\nu,R,h)$ will be the action taken by a customer with $\nu \leq \nu_o + 1$. The function $a^*(\nu,R,h)$ defines a policy for $\nu \leq \nu_o + 1$, all (R,h). We say this policy is optimal.

In the sequel we shall be concerned mostly with complete policies. We therefore extend the concept of optimality to optimal policies that are made complete.

DEFINITION (Optimality). Let $P:\{(\nu,R,h)\} \to A$ be a complete policy with balking limit ν_o. We say P is optimal if $P(\nu,R,h) = a^*(\nu,R,h)$ for $\nu \leq \nu_o + 1$, and $P(\nu,R,h) \in A^*(\nu,R,h)$ for $\nu > \nu_o + 1$.

If an optimal policy does not exist there must be some arriving customers who do not have an optimal action under the model specification (the inference $\{A^*(\nu,R,h)\}$ leaves too much uncertainty for determination of an optimal action). Still, each arrival has to make a choice. The decision will then be made on the basis of his assumptions of how other customers act. These behavioral assumptions should be in agreement with the inference that (ν,R,h)-customers select an action from $A^*(\nu,R,h)$. For stationary models like ours a strong case can be made (but we will not give the arguments) that all arriving customers accept identical behavioral assumptions and that these will take the form of a policy assumption. By that we mean the assumption that all other customers select an action as a function of (ν,R,h). Denoting such a policy assumption by P_1, we assume, for simplicity, that the domain of P_1 is $\{(\nu,R,h)\} = \{(\nu,R,h): \nu \leq N+1, (R,h) \in \Omega\}$, that is that P_1 is complete, and, usually, $P_1(\nu,R,h) \in A^*(\nu,R,h)$.

Given P_1, each (ν,R,h)-customer can, in principle, calculate the (steady-state) payoff, E(V), for each alternative action he may contemplate. Consequently, each arrival can determine an optimal or at least an ε-optimal action. If a (ν,R,h)-customer has an optimal action given P_1, we denote it by $P_2(\nu,R,h)$.

We are now ready to state our definition of a stable policy in the case of a complete policy assumption.

DEFINITION (Stability). Let $P_1: \{(\nu,R,h)\} \to A$ be a complete policy

(assumption) with balking limit ν_o. We say P_1 is a stable policy if (i) each (ν,R,h)-customer with $\nu \leq \nu_o + 1$ has an optimal action $P_2(\nu,R,h)$, given P_1, and (ii) $P_2(\nu,R,h) = P_1(\nu,R,h)$ for all $\nu \leq \nu_o + 1$.

Normally $P_1(\nu,R,h) \in A^*(\nu,R,h)$, but we do allow "unreasonable" behavioral assumptions in our definition of stability.

The justification for the terminology "stable" is clear. By definition, if a policy (assumption) is stable, all arriving customers will act in accordance with the assumed behavior. Thus there is no reason for future customers to change the policy assumption. Hence the observable customer actions define a policy that does not change over time. In short, a stable policy is self-sustaining.

Two complete policy assumptions, P_1' and P_1'', are said to be equivalent if the respective response policies, P_2' and P_2'', are well defined, have identical balking limit ν_o, and $P_2'(\nu,R,h) = P_2''(\nu,R,h)$ for $\nu \leq \nu_o + 1$.

The statement concerning the significance of the distinction between optimal and stable solutions to games, applies equally to optimal and stable policies in Models A, B and C.

3. Model A: A GI/M/s/N System with Balking Option

In Model A the service order is imposed as a rule, but each customer may join or balk as he desires. When assuming FIFO rule we shall consider the G/M/s/N system, a slightly generalized system, with general arrivals rather than general independent arrivals. Section 3.1 deals with FIFO rule. Section 3.2 deals with LIFO rule.

3.1. A G/M/s/N/FIFO ($N \leq \infty$) System with Balking Option

Of all cases the G/M/s/N/FIFO system with balking option is by far

the easiest one to analyze, almost to the point of being trivial. Still, for future reference, and completeness, we state our findings in Theorem 1 that covers both $N < \infty$ and $N = \infty$.

THEOREM 1. In a G/M/s/N/FIFO ($N \leq \infty$) system with balking option there exists an optimal policy as follows: if $-s+1 \leq \nu \leq 0$, join if $R-b(\nu,R,h)-h/\mu \geq 0$, and otherwise balk; if $1 \leq \nu \leq N$, join if $R-b(\nu,R,h) -h[s+\nu]/(s\mu) \geq 0$, and otherwise balk; if $\nu = N+1$, always balk.

PROOF. We omit the proof since it is straightforward. The policy clearly is optimal as every customer can make an optimal choice regardless of what others do. □

It is worth noting the arrival rate is irrelevant for the decision-making. In the special case that all customers are alike, i.e. (R,h) is a constant, the resulting policy is of the control limit type, that is, there is a number ν_o so that all customers will join as long as $\nu \leq \nu_o$, and balk if $\nu = \nu_o + 1$.

It is convenient to introduce special notation for Model A, taking advantage of the fact that it recognizes only two actions. A complete policy will be given as Δ: $\{(\nu,R,h)\} \to \{0,1\}$, where $\Delta(\nu,R,h) = 0$ if balking is specified, and $\Delta(\nu,R,h) = 1$ if joining is specified. Also let

$$f_\nu = \int_{\Delta(\nu,R,h)=1} dU(R,h) \tag{1}$$

$$g_\nu = \int_{\Delta(\nu,R,h)=0} dU(R,h) \tag{2}$$

We shall refer to f_ν as the joining fraction, and to g_ν as the balking fraction for ν-customers.

The same, or similar, notation will be employed for a policy assumption, and for the response policy that gives the optimal action under the policy assumption.

Analogously, let $\Delta^*(\nu,R,h)$ describe the reduced action sets as follows: $\Delta^*(\nu,R,h) = 0$ if balking is optimal ($A^*(\nu,R,h)=\{0\}$); $\Delta^*(\nu,R,h) = 1$ if joining is optimal ($A^*(\nu,R,h)=\{1\}$); $\Delta^*(\nu,R,h) = 1/2$ if the customer has insufficient information, as given by model, for a certain determination that either choice is optimal ($A^*(\nu,R,h)=\{0,1\}$).

3.2. A GI/M/s/N/LIFO ($N\leq\infty$) System with Balking Option

Because of significant differences between the cases $N < \infty$ and $N = \infty$ they will be treated separately, in Sections 3.2(b) and 3.2(c), respectively. First, however, we shall present some results that will be needed for calculation of the expected queueing time under LIFO rule.

3.2(a). The Mean Busy Period in an M/M/1/k System with Balking.

Recall that a busy period is the length of time from the epoch at which an arrival joins an empty system until the system again becomes empty. We are interested in the mean of the busy period in an M/M/1/k system in which a ν-customer joins with probability f_ν. Obviously, the mean busy period does not depend upon f_0. Lemma 1 gives us the mean busy period for all $k \geq 0$ and all f's.

LEMMA 1. Let $B_k(f_1,\ldots,f_k)$ denote the mean busy period in an M/M/1/k system with arrival rate λ, joining probability in state ν equal to f_ν ($\nu=1,\ldots,k$), and service rate μ. Then

$$B_k(f_1,\ldots,f_k) = \begin{cases} \dfrac{1}{\mu} & (k=0), \\ \dfrac{1}{\mu}\left[1 + \sum_{i=1}^{k}\left(\prod_{j=1}^{i} f_j\right)(\lambda/\mu)^i\right] & (0<k\leq\infty). \end{cases} \tag{3}$$

In particular, $B_k \equiv B_k(1,\ldots,1) = \frac{1}{\mu} \sum_{i=0}^{k} (\lambda/\mu)^i$ for $0 \leq k \leq \infty$.

PROOF. For $k = 0$ the result is obvious. Next consider $1 \leq k < \infty$. (Note, with $s = 1$ we have $\nu = n$.) Let $\lambda_i = f_i\lambda$ denote the rate of joining given state i. λ_0^{-1} then, is the mean idle period. Let p_0 be the steady-state probability that there are 0 customers in the system. With departure rate μ in all nonzero states it is known that $p_0^{-1} = 1 + \lambda_0/\mu + \lambda_0\lambda_1/\mu^2 + \ldots + \lambda_0\lambda_1\ldots\lambda_k/\mu^{k+1}$. Considering that busy and idle periods alternate in a renewal process, it is easy to see that $B_k(f_1,\ldots,f_k)/[B_k(f_1,\ldots,f_k) + \lambda_0^{-1}] = 1 - p_0$, or $B_k(f_1,\ldots,f_k) = \lambda_0^{-1}(p_0^{-1}-1)$. Insertion of the expression for p_0^{-1} and use of $\lambda_i = f_i\lambda$ yield Equation 3, for $0 \leq k < \infty$. In particular, we have $B_k \equiv B_k(1,\ldots,1) = (1/\mu) \sum_{i=0}^{k} (\lambda/\mu)^i$ for $0 \leq k < \infty$. The last equation is known to hold also for $k = \infty$, then usually written $B_\infty = (1/\mu)(1 - \lambda/\mu)^{-1}$ for $\lambda/\mu < 1$, and $B_\infty = \infty$ for $\lambda/\mu \geq 1$. Similarly, Equation 3 will hold for $k = \infty$. □

As a corollary to Lemma 1 we have:

COROLLARY 1. The mean queueing time of a joining ν-customer ($\nu \geq 1$) in an M/M/s/N/LIFO system in which ν-customers join with probability f_ν is

$$Q_\nu(f_{\nu+1}, f_{\nu+2}, \ldots, f_N) = \begin{cases} 1/(s\mu) & (\nu = N) \\ [1/(s\mu)]\left[1 + \sum_{i=1}^{N-\nu} \left(\prod_{j=1}^{i} f_{\nu+j} \right)(\lambda/(s\mu))^i\right] & (1 \leq \nu < N) \end{cases} \tag{4}$$

In particular, in a system with forced joining the mean queueing time equals $Q_\nu = [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i$ for $1 \leq \nu \leq N$.

PROOF. Observe, the queueing time of a ν-customer ($1 \leq \nu \leq N$) is distributed as the busy period in an M/M/1/N-ν system with arrival rate λ, joining probability $f_{\nu+j}$ in arrival state $j = 1,\ldots,N-\nu$, and service rate $s\mu$. An application of Lemma 1 then yields the corollary. (Note, Equation (4) is valid for $N = \infty$.) □

$Q_\nu(\ )$ does not depend on the subscript ν which is included for identification only. $Q_\nu(\ )$ has $N-\nu$ arguments. If $N-\nu = 0$ we may write $Q_\nu(\cdot)$. Also, we define $Q_\nu = Q_\nu(1,\ldots,1)$ and $Q_N = Q_N(\cdot)$. The notation $Q_\nu(f_{\nu+1},f_{\nu+2},\ldots,f_N)$ will be used also to designate the mean queueing time of a ν-customer who joins a GI/M/s/N/LIFO system in which ν-customers join with probability f_ν. This makes sense because, as one can easily verify, the queueing time of a joining ν-customer has the distribution of the busy period in the GI/M/1/N-ν system with the same interarrival distribution function G(t), joining probability $f_{\nu+j}$ in arrival state j, and service rate $s\mu$. However, we do not have a formula from which $Q_\nu(f_{\nu+1}, f_{\nu+2},\ldots,f_N)$ can be calculated in the case of a general interarrival distribution.

3.2(b). The Case $N < \infty$. We now turn our attention to a system with interarrival time distribution G(t), exponential service time distribution, finite waiting room, LIFO rule and balking option. Again, an optimal policy always exists as stated in the next theorem. The optimal decision for each (ν,R,h)-customer such that $1 \leq \nu \leq N$, is arrived at by backward induction.

THEOREM 2. In a GI/M/s/N/LIFO ($N<\infty$) system with balking option there exists an optimal policy as follows: if $-s+1 \leq \nu \leq 0$, join if $R-b(\nu,R,h)$

$-h/\mu \geq 0$, and otherwise balk; if $\nu = N+1$, always balk; if $1 \leq \nu \leq N$, join if $R-b(\nu,R,h)-h[1/\mu + Q_\nu(f_{\nu+1},f_{\nu+2},\ldots,f_N)] \geq 0$ $[\Delta(\nu,R,h)=1]$, and otherwise balk $[\Delta(\nu,R,h)=0]$, where f_ν is given by Equation (1) and $Q_\nu(f_{\nu+1},f_{\nu+2},\ldots,f_N)$ is the expected queueing time of a joining ν-customer, given by Equation (4) in the M/M/s/N/LIFO case. For $\nu = 1,\ldots,N$ the decision function $\Delta(\nu,R,h)$ is derived recursively through the steps: $Q_N(\cdot) = 1/(s\mu)$, $\Delta(N,R,h)$ for all $(R,h) \in \Omega$, f_N, $Q_{N-1}(f_N)$, $\Delta(N-1,R,h)$ for all $(R,h) \in \Omega$, $f_{N-1},\ldots,Q_1(f_2,f_3,\ldots,f_N)$, $\Delta(1,R,h)$ for all $(R,h) \in \Omega$, (f_1).

PROOF. For $-s+1 \leq \nu \leq 0$ and $\nu = N+1$ the optimal action is obvious. Consider an (N,R,h)-customer. Under LIFO rule his expected net gain by joining is $E(V) = R-b(N,R,h)-h[1/\mu + Q_N(\cdot)]$, with $Q_N(\cdot) = 1/(s\mu)$. Letting $\Delta(N,R,h) = 1$ if $E(V) \geq 0$ and $\Delta(N,R,h) = 0$ if $E(V) < 0$, the (N,R,h)-customer will, of course, join if $\Delta(N,R,h) = 1$, and otherwise balk. The joining probability for N-customers is f_N, given by (1), and all customers who need to know will know f_N. Now, knowing f_N an $(N-1,R,h)$-customer can determine his expected net gain by joining as $E(V) = R-b(N-1,R,h)$ $-h[1/\mu + Q_{N-1}(f_N)]$ where $Q_{N-1}(f_N)]$ is the mean queueing time of a joining $(N-1)$-customer, and then easily decide whether to join or balk. These decisions are represented by $\Delta(N-1,R,h)$ for all $(R,h) \in \Omega$, and this in turn forms the basis for calculation of f_{N-1}. Continuing this way we derive $\Delta(N,R,h)$, $\Delta(N-1,R,h),\ldots,\Delta(1,R,h)$ for all $(R,h) \in \Omega$. The policy $\Delta(\nu,R,h)$, now defined for every conceivable customer, is a complete policy, and by its derivation it is clearly optimal. □

Note, in the special case that all customers are alike, the resulting policy is of the control limit type, just as in the FIFO queue.

However, the LIFO queue with all customers alike exhibits some pathological characteristics. As $N \to \infty$ the optimal action by a ν-customer does not necessarily converge to either joining or balking. This result, easily verified, suggests that the GI/M/s/∞/LIFO balking model is a strange creature, as indeed it is.

3.2(c). The Case $N = \infty$. Perhaps one would expect the analysis to simplify when going from finite to infinite waiting room, but the opposite is true. Three questions will be dealt with here. First, we present a detailed derivation of the reduced action sets and the corresponding joining and balking fractions. Second, we show by example that an optimal policy need not exist, and that the number of stable policies may be greater than one. Third, we give necessary and sufficient conditions for the existence of an optimal policy.

As always, for $-s+1 \leq \nu \leq 0$, each ν-customer has an optimal action, and joining fractions, f_ν^*, and balking fractions, $g_\nu^* = 1-f_\nu^*$, are easily obtained by Equations (1) and (2). Now consider the decisions by customers arriving at a busy system.

The reduced action sets for $\nu \geq 1$ are derived in an inductive reasoning process as explained in Section 2.2. Assume simultaneous reduction. Let $\{A^i(\nu,R,h)\}$ denote the partially reduced (complete if i=1) action sets at stage i (prior to reduction). Let $\Delta^i(\nu,R,h) = 0, 1, 1/2$ according as $A^i(\nu,R,h) = \{0\}, \{1\}, \{0,1\}$. Let f_ν^i, g_ν^i denote the corresponding joining and balking fractions. As $A^1(\nu,R,h) = \{0,1\}$, we have $\Delta^1(\nu,R,h) = 1/2$ and $f_\nu^1 = g_\nu^1 = 0$. Now consider the result of the reduction at stage 1, namely $\{A^2(\nu,R,h)\}$ or, equivalently, $\{\Delta^2(\nu,R,h)\}$. Obviously, for $\nu \geq 1$,

$$\Delta^2(\nu,R,h) = \begin{cases} 0 \text{ if } R-b(\nu,R,h)-h[1/\mu + Q_\nu(0,\ldots)] < 0\ , \\ 1 \text{ if } R-b(\nu,R,h)-h[1/\mu + Q_\nu(1,1,1,\ldots)] \geq 0, \\ 1/2 \text{ otherwise}, \end{cases}$$

where $Q_\nu(0,\ldots) = 1/(s\mu)$, $Q_\nu(1,1,1,\ldots) = 1/(s\mu-\lambda)$ if $\lambda < s\mu$ and $Q_\nu(1,1,1,\ldots) = \infty$ if $\lambda \geq s\mu$. $\Delta^2(\nu,R,h) = 0$ means that balking is the best choice even under the most optimistic assumption that no higher-state customer will join; consequently the customer will balk. $\Delta^2(\nu,R,h) = 1$ means that joining is the best choice even under the most pessimistic assumption that all higher-state customers will join; consequently the customer will join. $\Delta^2(\nu,R,h) = 1/2$ means the customer does not have sufficient information at step 1 to determine an optimal decision. Clearly,

$$f_\nu^2 = \int_{\Delta^2(\nu,R,h)=1} dU(R,h), \qquad g_\nu^2 = \int_{\Delta^2(\nu,R,h)=0} dU(R,h).$$

At stage 2 each ν-customer will know $f_{\nu+j}^2$ and $g_{\nu+j}^2$ for all $j \geq 1$. The sequence $(f_{\nu+j}^2)_j$ will be his new most optimistic assumption of joining probabilities, and $(1-g_{\nu+j}^2)_j$ will be his new most pessimistic assumption. Hence,

$$\Delta^3(\nu,R,h) = \begin{cases} 0 \text{ if } R-b(\nu,R,h)-h[1/\mu + Q_\nu((f_{\nu+j}^2)_j)] < 0, \\ 1 \text{ if } R-b(\nu,R,h)-h[1/\mu + Q_\nu((1-g_{\nu+j}^2)_j)] \geq 0, \\ 1/2 \text{ otherwise}. \end{cases}$$

Thus, at stage 2, if a customer has an optimal choice it is indicated by

the value of $\Delta^3(\nu,R,h)$. The corresponding joining and balking fractions, f_ν^3 and g_ν^3, are calculated as before. A decision to join or balk clearly will not change from one stage to the next, so $f_\nu^1 \leq f_\nu^2 \leq f_\nu^3$ and $g_\nu^1 \leq g_\nu^2 \leq g_\nu^3$. Continuing in this manner we obtain increasing sequences $(f_\nu^i)_i$ and $(g_\nu^i)_i$ for each $\nu \geq 1$. The limits $f_\nu^* = \lim_i f_\nu^i$ and $g_\nu^* = \lim_i g_\nu^i$ always exist. They are the joining and balking fractions, respectively. The reduced action sets are given by $\Delta^*(\nu,R,h) = \lim_i \Delta^i(\nu,R,h)$, but may also be determined via calculation of $Q_\nu((1-g^*_{\nu+j})_j)$ and $Q_\nu((f^*_{\nu+j})_j)$, as above.

The assertion that multiple (nonequivalent) stable policies may exist will be proved by an example. Consider an M/M/s/∞/LIFO system with constant (R,h), $b(\nu,R,h) = 0$ for all ν, and parameters such that

$$R - h[1/\mu + [1/(s\mu)] \sum_{i=0}^{1} (\lambda/(s\mu))^i] \geq 0,$$

but

$$R - h[1/\mu + [1/(s\mu)] \sum_{i=0}^{2} (\lambda/(s\mu))^i] < 0.$$

In the present case no optimal policy exists, since for $\nu \leq 0$ clearly $\Delta(\nu,R,h) = 1$, but for $\nu \geq 1$ no ν-customer has an optimal action at stage 1 or later. However, there are exactly three nonequivalent stable policies, all with $\Delta(\nu,R,h) = 1$ for $\nu \leq 0$, but with different $(\Delta(\nu,R,h))_{\nu\geq 1}$ as follows: (a) (0,1,1,...), (b) (1,0,1,1,...), (c) (1,1,0,1,1,...). Here, "..." indicate arbitrary actions.

To see this, note that under a complete policy all (ν,R,h)-customers make the same decision. Since (R,h) is constant, then for all ν-customers, either $\Delta(\nu,R,h) = 0$ or $\Delta(\nu,R,h) = 1$. Now, clearly the policy (1,1,1,...) is not stable, because a 1-customer, who assumed that all

2-and 3-customers join, would, by the second inequality, be better off not joining. Thus stability requires that 1-, 2- or 3- customers balk. Considering the remaining possibilities it is readily verified that the three listed policies are the only stable policies.

The next theorem gives conditions for the existence of an optimal policy. It also specifies the optimal action for each arrival. As the theorem is based upon (f_ν^*) and (g_ν^*), the results may not be too useful.

THEOREM 3. In a GI/M/s/∞/LIFO system with balking option, an optimal policy exists if and only if (i) $g_\nu^* = 1$ for some $\nu \geq -s+1$, or (ii) $g_\nu^* < 1$ and $f_\nu^* + g_\nu^* = 1$ for all $\nu \geq -s+1$. In case (i) Theorem 2 applies. In case (ii) the policy is to join if either $R-b(\nu,R,h)-h/\mu \geq 0$ for $-s+1 \leq \nu \leq 0$, or if $R-b(\nu,R,h)-h[1/\mu + Q_\nu((f_{\nu+j}^*)_j)] \geq 0$ for $\nu \geq 1$, and to balk otherwise.

PROOF. First it will be shown the condition ((i) or (ii)) is necessary. Observe, the only two possibilities are: $g_\nu^* = 1$ for some ν, and $g_\nu^* < 1$ for all ν. When $g_\nu^* < 1$ for all ν, evidently we must have $f_\nu^* + g_\nu^* = 1$ for all ν in order to ensure that all arriving customers have an optimal choice as required for optimality. Thus the condition is necessary. Next we show the condition is sufficient. In case (i) there is, in effect, a limit on the queue length, and we can apply Theorem 2 which tells us that an optimal policy exists. In case (ii) the existence of an optimal policy follows directly from the definition of optimality. Hence the condition ((i) or (ii)) is necessary and sufficient for the existence of an optimal policy. In case (i) Theorem 2 gives the policy; N can be any ν such that $g_\nu^* = 1$. In case (ii) the policy follows from the discussion preceding the

theorem. □

More directly applicable is Theorem 4 which follows. It covers the special case where all customers are alike. That is, (R,h) = constant. To emphasize this fact we write $b(\nu)$ instead of $b(\nu,R,h)$.

THEOREM 4. In a GI/M/s/∞/LIFO system with a balking option and constant (R,h), an optimal policy exists if and only if

(i) $\lambda < s\mu$, $R-b(\nu)-h/\mu \geq 0$ for $\nu \leq 0$, and
$R-b(\nu)-h[1/\mu + 1/(s\mu-\lambda)] \geq 0$ for $\nu \geq 1$; or

(ii) $R-b(\nu)-h/\mu < 0$ for some $\nu \leq 0$ or
$R-b(\nu)-h[1/\mu + 1/(s\mu)] < 0$ for some $\nu \geq 1$; or

(iii) $\lambda < s\mu$, $R-b(\nu)-h/\mu \geq 0$ for $\nu \leq 0$, and
$R-b(\nu)-h[1/\mu + 1/(s\mu)] \geq 0$ for $\nu \geq 1$, and there exist $N \geq 0$ and k, $1 \leq k \leq \infty$, such that
$R-b(N+1+j)-h[1/\mu + 1/(s\mu-\lambda)] \geq 0$ $(j=1,k)$ and
$R-b(N+1)-h[1/\mu + Q_{N+1}(1,1,\ldots,1,0,\ldots)] < 0$ (with k leading 1's in Q_{N+1}).

In case (i) the policy is that every arrival joins. In cases (ii) and (iii) there is a balking limit $\nu_o < \infty$.

PROOF. Suppress the constant (R,h) and let $\Delta^i(\nu) \equiv \Delta^i(\nu,R,h)$, $\Delta^*(\nu) \equiv \Delta^*(\nu,R,h)$. In the present case an optimal policy, by its definition, will exist if and only if (a) $\Delta^*(\nu) = 1$ for all ν, or (b) $\Delta^*(\nu_o + 1) = 0$ for some $\nu_o \geq -s$, with $\Delta^*(\nu) = 1$ for all $\nu \leq \nu_o$. We must prove that (i), (ii) or (iii) is necessary and sufficient for (a) or (b). This we do by showing that (i) ⇒ (a), (ii) ⇒ (b), (iii) ⇒ (b), and that the

mutually exclusive conditions (i), (ii) and (iii) exhaust the ways (a) or (b) may occur.

At stage 1 we have $\Delta^1(\nu) = 1/2$ for all ν. Assume simultaneous reduction and consider all courses the reduction process may take. There are four distinct possibilities: (i) $\Delta^2(\nu) = 1$ for all ν; (ii) $\Delta^2(\nu) = 0$ for some ν; (iii) $\Delta^2(\nu) \neq 0$ for all ν, $\Delta^2(\nu) = 1$ for some ν, $\Delta^2(\nu) = 1/2$ for some ν, and $\Delta^3(\nu) = 0$ for some ν; (iv) $\Delta^2(\nu) \neq 0$ for all ν, $\Delta^2(\nu) = 1$ for some ν, $\Delta^2(\nu) = 1/2$ for some ν, and $\Delta^3(\nu) \neq 0$ for all ν. From the previous discussion of the reduction process for the GI/M/s/∞/LIFO system it follows that the above possibilities (i), (ii), (iii) are, respectively, equivalent to conditions (i), (ii), (iii) of the theorem.

Clearly, in case (i), $\Delta^*(\nu) = 1$ for all ν, so (i) $\Rightarrow$ (a). In case (ii), let N denote the smallest ν such that $\Delta^2(\nu+1) = 0$. If $N < 0$, set $\nu_o = N$. Then, obviously, $\Delta^*(\nu) = 1$ for $\nu \leq \nu_o$, and $\Delta^*(\nu_o+1) = 0$. If $N \geq 0$ we can, as in the proof of Theorem 2, show that there is a ν_o, $0 \leq \nu_o \leq N$, such that $\Delta^*(\nu) = 1$ for $\nu \leq \nu_o$, and $\Delta^*(\nu_o + 1) = 0$. Thus, (ii) $\Rightarrow$ (b). In case (iii), an optimal policy has not been identified at stage 1. However, at stage 2, for some $N \geq 0$, we find $\Delta^3(N+1) = 0$, implying as before the existence of a ν_o, $0 \leq \nu_o \leq N$, such that $\Delta^*(\nu) = 1$ for $\nu \leq \nu_o$, and $\Delta^*(\nu_o + 1) = 0$. (Notice, for the given $(\Delta^2(\nu))_\nu$ we can have $\Delta^3(N+1) = 0$ only if $\Delta^2(N+1) = 1/2$ and $\Delta^2(N+2) = 1$.) Hence, (iii) $\Rightarrow$ (b). On the other hand, condition (iv) can never lead to (a) or (b), since here $\Delta^2(\nu) = \Delta^3(\nu) = \Delta^*(\nu)$ for all ν. This concludes our proof that (i), (ii) or (iii) are necessary and sufficient for the existence of an optimal policy. $\square$

4. Model B: An M/M/s/N System with Priority Option

In Model B all arriving customers must join the system unless all waiting positions are occupied, but each customer who joins the queue may make any priority payment $b \in B$. The analysis will concern only the decision-making by customers who join the queue ($\nu=1,\ldots,N$) since other customers ($\nu \leq 0$ and $\nu=N+1$) have no choice to make. Note, the objective function of a customer joining the queue ($\nu=1,\ldots,N$) is $E(V) = R-b-hE(W)$. Since the customer will receive his reward R whatever choice $b \in B$ he makes, clearly, maximization of the expected net gain is equivalent to minimization of the expected cost $b+hE(W)$. Our main purpose in studying Model B is to state necessary and sufficient conditions for a customer policy to be stable and induce either FIFO or LIFO service order.

Here, as in the analysis of Model C we shall deal with stability based upon our dominance concept, strong dominance. As before, we speak of strong stability, but omit the word strong when no misunderstanding should occur. Besides we shall present results concerning stability of two different kinds, namely weak stability ($\geq$) and strict stability ($>$), discussed at length in Section 6. Our theorems cover all three kinds of stability. Proofs are given only for the case of strong stability. For convenience, we define a strongly (weakly, strictly) FIFO-stable policy as a policy which is strongly (weakly, strictly) stable and induces FIFO service order. A strongly (weakly, strictly) LIFO-stable policy is similarly defined.

The case $N = 1$ is trivial, so henceforth $N > 1$ will be assumed. Section 4.1 deals with the FIFO case, and Section 4.2 deals with the LIFO case.

4.1. FIFO-Stable Policies in an M/M/s/N ($N \leq \infty$) System with Priority Option

Because of significant differences between the cases $N < \infty$ and $N = \infty$ they will be discussed separately, in Section 4.1(b) and 4.1(c), respectively. We begin, in Section 4.1(a), by presenting a lemma which links FIFO service order to priority payments.

4.1(a). Priority Payment Policies That Induce FIFO Service Order.

LEMMA 2. Consider a G/M/s/N ($N \leq \infty$) system with forced joining and priority payments, where for $\nu = 1,\ldots,N$ the (ν,R,h)-customer pays $b(\nu,R,h)$. Then FIFO service order is induced if and only if $b(1,R,h) \geq b_o$ for all $(R,h) \in \Omega$, and $b(\nu,R,h) = b_o$ for all $(R,h) \in \Omega$, $\nu = 2,\ldots,N$, where b_o is a constant.

PROOF. Obviously, the condition is sufficient to induce FIFO service order. It will be shown the condition is also necessary. We start with the observation that it is necessary that $b(\nu,R',h') \geq b(\nu+1,r'',h'')$ for $\nu = 1,\ldots,N-1$ and all $(R',h'),(R'',h'') \in \Omega$. We will show that for $\nu \geq 2$ all ν-customers must pay the same, say $b(\nu)$. Suppose, on the contrary, that for some $\nu=\nu^* \geq 2$, $b(\nu^*,R,h)$ is not the same for all $(R,h) \in \Omega$. Imagine that $s+\nu^*$ customers arrive at an initially empty system, the last customer paying $b(\nu^*,R',h')$, and then a service completion occurs, followed by the arrival of another ν^*-customer paying $b(\nu^*,R'',h'')$. Now we may have $b(\nu^*,R',h') < b(\nu^*,R'',h'')$. This would violate the FIFO service order requirement. Hence, $b(\nu,R,h) = b(\nu)$ for $\nu = 2,\ldots,N$ is necessary. Thus we can assume $b(1,R,h) \geq b(2) \geq \ldots \geq b(N)$. Next, we will show $b(\nu) = b_o$ for $\nu = 2,\ldots,N$ is necessary. Suppose, on the contrary, that

for some $\nu = \nu^* \geq 2$, $b(\nu^*) > b(\nu^*+1)$. Imagine that $s+\nu^*+1$ customers arrive at an initially empty system, the last customer paying $b(\nu^*+1)$, and then two service completions take place, followed by the arrival of another customer, who is a ν^*-customer paying $b(\nu^*)$. Since the last arrival pays $b(\nu^*)$, which is more than the $b(\nu^*+1)$ paid by the preceding arrival, still in the queue, the FIFO service order requirement is not met. We conclude $b(1,R,h) \geq b(2) = \ldots = b(N)$. Hence, the condition is also necessary as claimed. □

4.1(b). The Case $N < \infty$.

THEOREM 5. In an M/M/s/N ($2 \leq N < \infty$) system with forced joining and priority option, the policy of paying $b=b(\nu,R,h) \in B$ in arrival states $\nu = 1,\ldots,N$ is FIFO-stable if and only if

(i) $b(\nu,R,h) = b_o \in B \quad (\nu=1,\ldots,N)$, and

(ii) if $b_o > \inf b$, then $b_o \underset{(=)}{\leq} \bar{b} \equiv \inf b + \inf h\, [1/(s\mu)] \sum_{i=1}^{N-1} (\frac{\lambda}{s\mu})^i$, and

(iii) if $b_o < \sup b$, then $b_o \underset{(=)}{\leq} \inf \{b:\ b\in B,\ b>b_o\} - \sup h\, [1/(s\mu)](N-1)$.

In (ii) we have $<$ only in case of strict or strong FIFO stability and only if $\inf b \in B$ and $\inf h \in H$. In (iii) we have $<$ only in case of strict FIFO-stability and only if $\inf \{b:\ b\in B,\ b>b_o\} \in B$ and $\sup h \in H$. A FIFO-stable policy need not exist for any $\lambda/(s\mu)$. If $\lambda/(s\mu) < 1$, and a FIFO-stable policy exists, then it is unique, given by $b(\nu,R,h) = b_o$, with $b_o \in B$ being the greatest value satisfying (ii). If $\lambda/(s\mu) \geq 1$, a FIFO-stable policy is not necessarily unique.

PROOF. We shall prove the theorem only for the case of strong FIFO-stability, omitting the descriptor "strong" throughout. Proofs for the cases of weak and strict FIFO-stability are nearly identical to the one

given here. We begin by showing conditions (i), (ii), (iii) are necessary for FIFO-stability. First, by Lemma 2, FIFO service order requires $b(\nu,R,h) = b_o \in B$ for $\nu = 2,\ldots,N$, all $(R,h) \in \Omega$ and $b(1,R,h) \geq b_o$ for all $(R,h) \in \Omega$. Further, stability requires $b(1,R,h) = b_o$, since a 1-customer can only lose by paying more. Hence, (i) is necessary.

Second, stability requires that if a customer should make a payment which is lower than the policy payment b_o (assumes $b_o > \inf b$), then his expected cost must exceed the cost associated with b_o. The expected cost for a (ν,R,h)-customer who pays b_o like everyone else is $b_o + h[1/\mu + \nu/(s\mu)]$. On the other hand, if he pays $b < b_o$ while others pay b_o, his expected cost will be $b + h[1/\mu + \sum_{k=N-\nu}^{N-1} [1/(s\mu)] \sum_{i=0}^{k} (\lambda/(s\mu))^i]$. To see this, observe that in the latter case our customer will be the last one served, and his expected queueing time therefore is the sum of ν simple busy periods with means $[1/(s\mu)] \sum_{i=0}^{k} (\lambda/(s\mu))^i$ for $k = N-\nu,\ldots,N-1$, respectively, as one can see by serving all other customers LIFO-wise (which has no effect upon our customer's queueing time distribution) and applying the formula for Q_ν in Corollary 1, replacing N by N-1. It follows that stability against lower payment requires

$$b_o + h[1/\mu + \nu/(s\mu)] < b + h[1/\mu + \sum_{k=N-\nu}^{N-1} [1/(s\mu)] \sum_{i=0}^{k} (\lambda/(s\mu))^i] \qquad (6)$$

all $b<b_o$, $b\in B$,

all $h\in H$, $\nu=1,\ldots,N$.

Rewriting (6) we obtain

$$b_o < b + h[\sum_{k=N-\nu}^{N-1} [1/(s\mu)](\sum_{i=0}^{k} (\lambda/(s\mu))^i - 1)], \qquad (7)$$

all $b<b_o$, $b\in B$,

all $h\in H$, $\nu=1,\ldots,N$.

$\sum_{i=0}^{k} (\lambda/(s\mu))^i - 1$ is 0 for $k = 0$ and positive for $k \geq 1$, so the double sum of (7) is positive and attains its minimum for $\nu = 1$. We conclude that stability against lower payment requires $b_o < \inf b + \inf h\,[1/(s\mu)] \cdot \left(\sum_{i=0}^{N-1} (\lambda/(s\mu))^i - 1\right)$, if $\inf b \in B$ and $\inf h \in H$, and $b_o \leq \inf b + \ldots$ otherwise. Hence, (ii) is necessary.

Third, stability requires that if a customer should make a payment which is higher than the policy payment b_o (assumes $b_o < \sup b$), then his expected cost must exceed or equal the cost associated with b_o. The expected cost for a (ν,R,h)-customer paying $b > b_o$ is $b + h[1/\mu + 1/(s\mu)]$, since any payment higher than b_o will guarantee a customer he will be served next, if all others pay b_o. It follows that stability against higher payment requires

$$b_o + h[1/\mu + \nu/(s\mu)] \leq b + h[1/\mu + 1/(s\mu)] \qquad \begin{array}{l}\text{all } b>b_o,\ b\in B,\\ \text{all } h\in H,\ \nu=1,\ldots,N\end{array} \tag{8}$$

Rewriting (8) we obtain

$$b_o \leq b - h[1/(s\mu)](\nu-1) \qquad \begin{array}{l}\text{all } b>b_o,\ b\in B,\\ \text{all } h\in H,\ \nu=1,\ldots,N\end{array} \tag{9}$$

The conclusion, that (iii) is necessary, follows easily.

Conditions (i), (ii), (iii) are also sufficient. Obviously (i) effects the desired service order. Also, for fixed priority payment b_o, (ii) $\Rightarrow$ (6), (iii) $\Rightarrow$ (8), and (6) and (8) together guarantee stability against all payments other than b_o, so conditions (i), (ii), and (iii) are indeed sufficient for FIFO-stability.

Observe, (ii) places an upper limit $\bar{b}$ on b_o, while (iii) places a lower limit on the difference between b_o and the next higher payment in B.

Thus, whatever the values of $\lambda/(s\mu)$ and $N > 1$, if the elements of B are close enough, no FIFO-stable policy can exist except possibly $b_o = \sup b$. Now suppose $\lambda/(s\mu) < 1$. Assume $b_o \in B$ satisfies (ii). Then a $b_o^* < b_o$ with $b_o^* \in B$ also satisfies (ii), but it cannot satisfy (iii) since

$$\begin{aligned}
\inf \{b: b\in B,\ b>b_o^*\} - b_o^* &\leq b_o - \inf b \\
&\leq \inf h\ [1/(s\mu)] \sum_{i=1}^{N-1} (\lambda/(s\mu))^i \quad \text{[by (ii)]} \\
&\leq \sup h\ [1/(s\mu)] \sum_{i=1}^{N-1} (\lambda/(s\mu))^i \\
&< \sup h\ [1/(s\mu)](N-1) \quad [\sup h > 0,\ N>1]
\end{aligned}$$

is in violation of condition (iii). We conclude that $b(\nu,R,h) = b_o$ can be a FIFO-stable policy only if b_o is the maximum of those elements of B that satisfy (ii), including an $\inf b \in B$.

Finally, it is not difficult to demonstrate that, if $\lambda/(s\mu) \geq 1$, then more than one FIFO-stable policy may exist. For a counterexample with $\lambda/(s\mu) > 1$, let $0 < \inf h < \sup h < \infty$, select s, μ and $N>1$ so that $\sup h\ [1/(s\mu)](N-1) = 1/2$, and let $B = \{1,2,\ldots,10\}$. In addition, choose λ sufficiently large so that $\lambda/(s\mu) > 1$ and $\bar{b} > \sup b = 10$. Then it can be easily verified that any $b_o \in B$ satisfies (i), (ii), and (iii). For a counterexample with $\lambda/(s\mu) = 1$, let $\inf h = \sup h = h_o > 0$, $\inf b \in B$, $\bar{b} \in B$, $B = \{\inf b, \bar{b}\}$. Then it can be easily verified that both $b_o = \inf b$ and $b_o = \bar{b}$ satisfy (i), (ii), and (iii).

4.1(c). The Case $N = \infty$.

THEOREM 6. In an $M/M/s/\infty$ system with forced joining and priority option the policy of paying $b = b(\nu,R,h) \in B$ in arrival states $\nu = 1,2,\ldots$ is

FIFO-stable if and only if

(i) $b(\nu,R,h) = \sup b \in B \quad (\nu=1,2,\ldots)$, and

(ii) $\lambda/(s\mu) \geq 1$, or, $\lambda/(s\mu) < 1$ and $\sup b - \inf b \underset{(=)}{\leq} \inf h \cdot \frac{\lambda}{s\mu} \frac{1}{s\mu-\lambda}$.

In (ii) we have $<$ in the last inequality only in case of strict or strong FIFO-Stability and only if $\inf b \in B$ and $\inf h \in H$.

PROOF. As before we shall prove the theorem only for the case of strong FIFO-stability, and we suppress the word "strong." We begin by showing that conditions (i), (ii), are necessary for FIFO-stability. First, by Lemma 2 and the stability requirement we deduce $b(\nu,R,h) = b_o \in B$. Replacing (6) as the requirement for stability against lower payment we have

$$b_o + h[1/\mu + \nu/(s\mu)] < b + h[1/\mu + (\nu/(s\mu)) \sum_{i=0}^{\infty} (\lambda/(s\mu))^i] \qquad \begin{matrix}\text{all } b<b_o,\ b\in B,\\ \text{all } h\in H,\ \nu=1,2,\ldots\end{matrix} \qquad (6')$$

while (8) continues to be the requirement for stability against higher payment. Now however, no $b_o < \sup b$ can satisfy (8) for all ν, so we must have $\sup b \in B$ and $b_o = \sup b$. Hence (i) is necessary. Since no $b > \sup b$ exists, (i) in effect eliminates (8), and we are left with (6') which, using $b_o = \sup b$, can be written as

$$\sup b < b + h(\nu/(s\mu))\left(\sum_{i=0}^{\infty} (\lambda/(s\mu))^i - 1\right) \qquad \begin{matrix}\text{all } b<\sup b,\ b\in B,\\ \text{all } h\in H,\ \nu=1,2,\ldots\end{matrix} \qquad (7')$$

It is easily shown that (7') $\Rightarrow$ (ii). Hence, (i) and (ii) are necessary.

Next we show (i) and (ii) are sufficient conditions. Clearly, (i) implies FIFO service order. In addition (i) $\Rightarrow$ (8) (for $b_o = \sup b$). Also, (ii) $\Rightarrow$ (7') $\Rightarrow$ (6') (for $b_o = \sup b$). Now, the satisfaction of (6') and (8) for $b_o = \sup b$ guarantees stability of the policy $b(\nu,R,h) = \sup b$.

Hence, (i) and (ii) ensure FIFO-stability. Our proof is complete. Note that here, if a FIFO-stable policy exists, then it is unique. □

4.2. LIFO-Stable Policies in an M/M/s/N ($N \leq \infty$) System with Priority Option

Again there are significant differences between the cases $N < \infty$ and $N = \infty$ and we choose to discuss the two cases separately, in Section 4.2(c) and 4.2(d), respectively. First, however, we present a lemma, in Section 4.2(a), that links LIFO service order to priority payments. Also, in Section 4.2(b), we calculate certain expected queueing times needed in testing a policy for stability. The final Section 4.2(e) deals with the design of the priority payment set B.

4.2(a). Priority Payment Policies That Induce LIFO Service Order.

LEMMA 3. Consider a G/M/s/N ($N \leq \infty$) system with forced joining and priority payments, where for $\nu = 1,\ldots,N$ the (ν,R,h)-customer pays $b(\nu,R,h)$. Then LIFO service order is induced if and only if $b(\nu,R',h') < b(\nu+1,R'',h'')$ for all $(R',h') \in \Omega$, all $(R'',h'') \in \Omega$, $\nu = 1,\ldots,N-1$.

PROOF. Obvious. □

4.2(b). Some Expected Queueing Times. Consider an M/M/s/N ($N \leq \infty$) system with forced joining and priority option. In this subsection it will be assumed that $B = \{b_j\}_{1 \leq j \leq N}$ when N is finite and $B = \{b_j\}_{j \geq 1}$ when $N = \infty$, where (b_j) is a strictly increasing sequence of nonnegative numbers. We shall consider the policy of paying $b = b_\nu$ in state $\nu \geq 1$. Observe that, by Lemma 3, the given policy will result in LIFO service order. In order to test this policy for stability we need to know the expected queueing time for each arrival state $\nu \geq 1$, for any choice of priority payment from

the set $\{b_j\}$, on the condition that all other customers make the policy payment.

Define $q_{\nu\delta}$ $(1 \leq \nu \leq N,\ 1 \leq \delta \leq N)$ as the expected queueing time of a ν-customer who decides to pay b_δ, like a regular δ-customer, rather than b_ν, given that all others make the policy payment when joining the queue. By Corollary 1, $q_{\nu\nu} = Q_\nu = [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i$ for $1 \leq \nu \leq N$. In other cases the expected queueing time is given by one of the following expressions:

$$q_{\nu\delta} = \sum_{j=\delta}^{\nu} Q_j = [1/(s\mu)] \sum_{j=\delta}^{\nu} \sum_{i=0}^{N-j} (\lambda/(s\mu))^i \qquad (1 \leq \delta \leq \nu \leq N), \tag{10}$$

$$q_{\nu\delta} = 1/(s\mu) + \sum_{j=1}^{N-\delta} (\lambda/(\lambda+s\mu))^{\delta-\nu+j}\, Q_{\delta+j-1} \tag{11}$$

$$= 1/(s\mu) + [1/(s\mu)] \sum_{j=1}^{N-\delta} \{(\lambda/(\lambda+s\mu))^{\delta-\nu+j} \sum_{i=0}^{N-\delta-j+1} (\lambda/(s\mu))^i\}$$

$$(1 \leq \nu \leq \delta \leq N-1),$$

$$q_{\nu N} = 1/(s\mu) \qquad (1 \leq \nu \leq N). \tag{12}$$

First, let $1 \leq \delta < \nu \leq N$. In this case our ν-customer, C_1, is paying less than the policy payment. Consequently $\nu-\delta$ customers in the queue when C_1 arrives will have higher priority than C_1. Call these customers $C_2, C_3, \ldots, C_{\nu-\delta+1}$, arranged by increasing priority. C_1's queueing time is the sum of $\nu-\delta+1$ busy periods associated with $C_{\nu-\delta+1}, \ldots, C_2, C_1$, in that order. Let C_j generate busy period no. j. Then the first busy period, no. $\nu-\delta+1$, is the time from C_1's arrival until $C_{\nu-\delta+1}$ goes into service, and for $1 \leq j \leq \nu-\delta$ busy period no. j is the time from C_j becomes the highest

priority customer (by C_{j+1}'s removal from the queue) until C_j goes into service. The mean of each busy period is given by Corollary 1. (We have been careful to take into account C_1's presence in the queue.) Equation (10) is simply the result of the summation of the $\nu-\delta+1$ mean busy periods. Note (10) is valid also for $\delta = \nu$.

Next, let $1 \leq \nu \leq \delta \leq N-1$. Observe, we allow $\delta = \nu$. In this case C_1's payment gives him a higher priority than any customer in the queue at his arrival, but it is not high enough to guarantee that he will be served first. We may conceive of C_1's queueing time as being composed of two time intervals. Let I_1, I_2 denote these intervals as well as their respective length. I_1 is the time from C_1's arrival until the next service completion, and I_2 is the remainder of C_1's queueing time, if any. Clearly, $E(I_1) = 1/(s\mu)$. If at the end of I_1 C_1 is still the customer with the highest priority, then he will enter service immediately, and $I_2 = 0$. Otherwise C_1 must wait until all higher priority customers as well as himself are cleared from the queue. It is seen that, if not equal to zero, I_2 is the sum of busy periods of the kind just encountered. If $\delta-\nu$ or less customers arrive during I_1, clearly $I_2 = 0$. If more than $\delta-\nu$ customers should arrive during I_1, denote by C_2, C_3,..., C_{k+1} the k joining customers who obtain higher priority than C_1, where $1 \leq k \leq N-\delta$. At the end of I_1 a service completion occurs and C_{k+1} will go into service, leaving behind in the queue C_1, C_2,..., C_k besides lower priority customers. Now, at the start of I_2 a C_j $(1 \leq j \leq N-\delta)$ will be present in the queue if and only if at least $\delta-\nu+j$ customers arrive during I_1. The probability thereof is $(\lambda/(\lambda+s\mu))^{\delta-\nu+j}$. Also, the mean busy period associated with C_j, if present, is $Q_{\delta+j-1}$. Hence $E(I_2) = \sum_{j=1}^{N-\delta} (\lambda/(\lambda+s\mu))^{\delta-\nu+j} Q_{\delta+j-1}$. Equation (11) results

by adding $E(I_1)$ and $E(I_2)$.

Finally, let $1 \leq \nu \leq N$, $\delta = N$. In this case C_1 buys the highest priority and surely will be served first. Hence his mean queueing time is $1/(s\mu)$.

4.2(c). The Case $N < \infty$. The next theorem gives necessary and sufficient conditions that a priority payment policy $b = b(\nu,R,h) \in B$ is LIFO-stable. Notice, if a LIFO-stable policy exists, then it is unique.

THEOREM 7. In an M/M/s/N ($2 \leq N < \infty$) system with forced joining and priority option the policy of paying $b = b(\nu,R,h) \in B$ in arrival states $\nu = 1,\ldots,N$ is LIFO-stable if and only if

(i) B contains N smallest values $b_1 < b_2 < \ldots < b_N$, and

(ii) $b(\nu,R,h) = b_\nu$ $(\nu=1,\ldots,N)$, and

(iii) $\sup h \frac{\lambda}{\lambda+s\mu} \frac{1}{s\mu} \sum_{i=0}^{N-\nu} (\frac{\lambda}{s\mu})^i \; (\leq) \; b_{\nu+1} - b_\nu \; (\leq) \inf h \frac{1}{s\mu} \sum_{i=0}^{N-\nu} (\frac{\lambda}{s\mu})^i$ $(\nu=1,\ldots,N-1)$.

In (iii) we have $<$ in the first inequality only in case of strict LIFO-stability and only if $\sup h \in H$; and we have $<$ in the second inequality only in case of strict or strong LIFO-stability and only if $\inf h \in H$.

PROOF. We present a proof only for the case of strong LIFO-stability. We begin by showing that (i), (ii), and (iii) are necessary conditions. First, we show (i) and (ii) are necessary. By Lemma 3 LIFO service order requires $b(\nu,R',h') < b(\nu+1,R'',h'')$ for all $(R',h') \in \Omega$, all $(R'',h'') \in \Omega$, $\nu = 1,\ldots,N-1$. We now apply the stability requirement in a limited way. Let each customer assume a payment policy $b = b(\nu,R,h)$ that meets the above requirement for LIFO service order. A moment's reflection will show that the policy cannot be stable unless B contains a smallest value b_1 and $b(1,R,h) = b_1$ for all $(R,h) \in \Omega$, because otherwise some 1-customer can

effect a saving on priority payment without affecting the service order under the policy. Suppose therefore $b(1,R,h) = b_1$ for all $(R,h) \in \Omega$, and consider once more the question of stability. Now we find that the policy cannot be stable unless B contains a 2nd smallest value b_2 and $b(2,R,h) = b_2$ for all $(R,h) \in \Omega$. By induction we prove conditions (i) and (ii) are necessary.

Next we show (iii) is necessary, employing conditions (i) and (ii) that we have already shown are necessary. Thus consider a policy $b = b(\nu,R,h) = b_\nu$ for $\nu = 1,\ldots,N$ where $b_1,b_2,\ldots,b_N$ is a strictly increasing sequence of payments that are the N smallest in the set B.

Stability implies stability against lower payment. Thus it is required that should a ν-customer pay less than b_ν, while assuming that everyone else makes the policy payment, then his expected cost will exceed the cost associated with b_ν. That is, stability requires that $b_\nu + hq_{\nu\nu} < b_\delta + hq_{\nu\delta}$ for $1 \leqq \delta < \nu \leqq N$ and all $h \in H$, with $q_{\nu\nu}$ and $q_{\nu\delta}$ as defined in Section 4.2(b). It follows that

$$b_\nu - b_\delta \ (\leqq) \ \inf h\,(q_{\nu\delta} - q_{\nu\nu}) \qquad (1 \leqq \delta < \nu \leqq N),$$

with $\leqq$ if $\inf h \notin H$, and $<$ if $\inf h \in H$. By (10) $q_{\nu\delta} - q_{\nu\nu} = \sum_{j=\delta}^{\nu} Q_j - Q_\nu = \sum_{j=\delta}^{\nu-1} Q_j$. Hence the above condition for stability against lower payment can be stated as

$$b_\nu - b_\delta \ (\leqq) \ \inf h \sum_{j=\delta}^{\nu-1} Q_j \qquad (1 \leqq \delta < \nu \leqq N). \tag{13}$$

We will show (13) $\Leftrightarrow$ (14), where

$$b_{\nu+1} - b_\nu \ (\leqq) \ \inf h\, Q_\nu \qquad (1 \leqq \nu \leqq N-1), \tag{14}$$

i.e. stability against lower payment is ensured by stability against the next lower payment. Obviously, (13) $\Rightarrow$ (14), as seen by letting $\delta = \nu-1$ in (13). To prove (14) $\Rightarrow$ (13) let $1 \leq \delta < \nu \leq N$. Then

$$b_\nu - b_\delta = \sum_{j=\delta}^{\nu-1} (b_{j+1} - b_j)$$

$$(\leq) \quad \inf\ h \sum_{j=\delta}^{\nu-1} Q_j \qquad \text{[by (14)]}.$$

This proves (13) $\Leftrightarrow$ (14). Therefore (14) is the necessary condition sought. Substitution of $Q_\nu = [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i$ into (14) yields the second set of inequalities of condition (iii).

Similarly, stability against higher payment requires

$$b_\delta - b_\nu \geq \sup\ h\ (q_{\nu\nu} - q_{\nu\delta}) \qquad (1 \leq \nu < \delta \leq N).$$

Observe, we need not consider payments higher than b_N since a customer making this payment will always be the first one served. By (11) and (12), $q_{\nu\nu} - q_{\nu\delta} = \sum_{j=\nu}^{\delta-1} (q_{\nu j} - q_{\nu j+1}) = \sum_{j=\nu}^{\delta-1} (\lambda/(\lambda+s\mu))^{j-\nu+1} Q_j$ for $1 \leq \nu < \delta \leq N$. Hence, the above condition for stability against higher payment can be stated as

$$b_\delta - b_\nu \geq \sup\ h \sum_{j=\nu}^{\delta-1} \left(\frac{\lambda}{\lambda+s\mu}\right)^{j-\nu+1} Q_j \qquad (1 \leq \nu < \delta \leq N). \tag{15}$$

We will show (15) $\Leftrightarrow$ (16), where

$$b_{\nu+1} - b_\nu \geq \sup\ h\ \frac{\lambda}{\lambda+s\mu}\ Q_\nu \qquad (1 \leq \nu \leq N-1), \tag{16}$$

i.e. stability against higher payment is ensured by stability against the

next higher payment. Obviously, (15) $\Rightarrow$ (16), as seen by letting $\delta = \nu+1$ in (15). To prove (16) $\Rightarrow$ (15) let $1 \leq \nu < \delta \leq N$. Then

$$b_\delta - b_\nu = \sum_{j=\nu}^{\delta-1} (b_{j+1} - b_j)$$

$$\geq \sup h \sum_{j=\nu}^{\delta-1} \frac{\lambda}{\lambda+s\mu} Q_j \qquad \text{[by (16)]}$$

$$\geq \sup h \sum_{j=\nu}^{\delta-1} \left(\frac{\lambda}{\lambda+s\mu}\right)^{j-\nu+1} Q_j .$$

This proves (15) $\Leftrightarrow$ (16). Therefore (16) is the necessary condition sought. Substitution of $Q_\nu = [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i$ into (16) yields the first set of inequalities of condition (iii). This completes the proof that (iii) is a necessary conditon for LIFO-stability.

It remains to demonstrate that (i), (ii), (iii) together are sufficient conditions for LIFO-stability. By Lemma 3, (i) and (ii) guarantee LIFO service order. (iii) is the same as (14) and (16). Backtracking we find (14) $\Rightarrow$ (13) $\Rightarrow \langle\ b_\nu - b_\delta \ (\leq)\ \inf h\ (q_{\nu\delta} - q_{\nu\nu});\ 1 \leq \delta < \nu \leq N\ \rangle$ $\Rightarrow \langle\ b_\nu + hq_{\nu\nu} < b_\delta + hq_{\nu\delta};\ 1 \leq \delta < \nu \leq N,\ h \in H\ \rangle$. This establishes stability against lower payment. Similarly, we can prove stability against higher payment. Thus, (i), (ii), (iii) are necessary and sufficient conditions for LIFO-stability. □

4.2(d). The Case $N = \infty$. We shall not go into a detailed discussion of this particular case, since it would be largely a repetition of the discussion concerning the case $N < \infty$. The truth of Theorem 8 should become apparent by an examination of the effect of letting $N \to \infty$ in Theorem 7.

Previously Balachandran [1972, p. 321] has dealt with the case $N = \infty$ for $s = 1$, h = constant, and weak stability, obtaining conditions which are in agreement with ours, although stated differently.

THEOREM 8. In an M/M/s/∞ system with forced joining and priority option the policy of paying $b = b(\nu,R,h) \in B$ in arrival states $\nu = 1,2,\ldots$ is LIFO-stable if and only if

(0) $\lambda/(s\mu) < 1$, and

(i) B is a countable set $\{b_j\}_{j\geq 1}$ where (b_j) is a strictly increasing sequence, and

(ii) $b(\nu,R,h) = b_\nu$ $(\nu=1,2,\ldots)$, and

(iii) $\sup h \frac{\lambda}{\lambda+s\mu} \frac{1}{s\mu-\lambda} \overset{\leq}{(=)} b_{\nu+1} - b_\nu \overset{\leq}{(=)} \inf h \frac{1}{s\mu-\lambda}$ $\quad (\nu=1,2,\ldots)$.

In (iii) we have $<$ in the first inequality only in case of strict LIFO-stability and only if $\sup h \in H$; and we have $<$ in the second inequality only in case of strict or strong LIFO-stability and only if $\inf h \in H$.

4.2(e). Design of Priority Payment Set B. When the waiting loss factor h is the same for all customers it is always possible to construct a priority payment set B such that a corresponding LIFO-stable policy exists (in case $N = \infty$ we must have $\lambda/(s\mu) < 1$). B is constructed as follows: $b_1 \geq 0$ is arbitrary, b_2 is selected in any way such that the two inequalities of (iii) involving $b_2 - b_1$ are satisfied, and so on. If, on the other hand, h is a variable, it is not always possible to construct a set B with the desired properties. Corollary 2 states conditions for the existence of such a set.

COROLLARY 2. In an M/M/s/N $(2\leq N\leq\infty)$ system with forced joining and priority

option, where $\lambda/(s\mu) < 1$ if $N = \infty$, a priority payment set B to which there corresponds a LIFO-stable policy can be constructed if and only if $\sup h \underset{(=)}{\leq} [1+(s\mu/\lambda)] \inf h$, where we have $<$ only in case of strict or strong LIFO-stability and only if $\inf h \in H$ (strict and strong) or $\sup h \in H$ (strict).

PROOF. The corollary will be proved only for the case of strong LIFO-stability. We first prove it for $N < \infty$. Recall, Theorem 7 gives necessary and sufficient conditions for LIFO-stability. First we prove necessity. Assume a LIFO-stable policy exists. Then (i), (ii), (iii) of Theorem 7 are satisfied. Letting $c_\nu = [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i$, we have: (iii) $\Rightarrow \sup h\, [\lambda/(\lambda+s\mu)]\, c_\nu \underset{(=)}{\leq} \inf h\, c_\nu$ $(\nu=1,\ldots,N-1) \Rightarrow \sup h \underset{(=)}{\leq} [1+(s\mu/\lambda)] \inf h$. Conversely, clearly the satisfaction of the inequality allows the construction of a priority payment set B that will meet conditions (i) and (iii); and (ii) can always be satisfied. Hence, the condition $\sup h \underset{(=)}{\leq} [1+(s\mu/\lambda)] \inf h$ is necessary and sufficient. For $N = \infty$, the corollary is proved in a similar fashion using Theorem 8. □

5. Model C: An M/M/s/N System with Balking and Priority Option

In Model C each customer is free to join or balk, and if he joins the queue he can choose any priority payment $b \in B$. As might be expected the analysis of Model C is considerably more complicated than the analysis of Model A and Model B, that have just one option. A complete analysis will not be attempted. In Section 5.1 we indicate how to perform a test for FIFO-stability or LIFO-stability of a complete policy. In Section 5.2 we give necessary and sufficient conditions for stability of one particular, simple policy. As in the previous section the analysis focuses on

strongly stable policies, but we deal also with strict and weak stability.

5.1. Test for FIFO- or LIFO-Stability of a Complete Policy

Consider a complete policy with balking limit ν_o. Let the choice on balking/joining be quantified by $\Delta(\nu,R,h)$, with value 0 in case of balking and 1 in case of joining. The policy specifies $\Delta(\nu,R,h)$ for all $(R,h) \in \Omega$ and $\nu = -s+1,\ldots,N$ as well as $b(\nu,R,h) \in B$ whenever $\Delta(\nu,R,h) = 1$ and $\nu \geq 1$. We want to know whether the given policy is FIFO-stable, or LIFO-stable, for the M/M/s/N ($N \leq \infty$) system under study. Assume $N > 1$.

We start with conditions that are easy to prove and easy to apply. FIFO-stability requires that, for $\nu = 1,\ldots,\nu_o$, $\Delta(\nu,R,h) = 1 \Rightarrow b(\nu,R,h) = b_o \in B$, where b_o is a constant. LIFO-stability requires, that, for $\nu = 1,\ldots,\nu_o$, $\Delta(\nu,R,h) = 1 \Rightarrow b(\nu,R,h) = b_\nu \in B$, where (b_ν) is a strictly increasing sequence and $b_1,\ldots, b_{\nu_o}$ are the smallest elements in B. These results correspond to those obtained for Model B, and are proved in the same way.

Assuming that the above conditions are met, the next step of the test is to establish stability of the policy decision for every customer who might arrive under the policy, and who acts on the assumption that other customers follow the policy. There are, in general, three categories of customers to consider: (a) $\Delta(\nu,R,h) = 1$ with $\nu \leq \nu_o$; the customer must not gain by balking, or, for $1 \leq \nu \leq \nu_o$, by making a priority payment other than b_o (FIFO case), or b_ν (LIFO case), (b) $\Delta(\nu,R,h) = 0$ with $\nu \leq \nu_o$; if $\nu \leq 0$ the customer must not gain by joining, and if $1 \leq \nu \leq \nu_o$ the customer must not gain by joining, whether he makes the "policy payment" b_o (FIFO case) or b_ν (LIFO case), or makes any other payment, (c) $\Delta(\nu,R,h) = 0$ with $\nu = \nu_o + 1$; the customer must not gain by joining, for

any payment $b \in B$; generally, for this category of customers the specification of actions for $\nu > \nu_o + 1$ is needed, unless $\nu_o = N$.

In principle the statements made above can be translated into quantitative necessary and sufficient conditions for FIFO- or LIFO-stability. However, in some cases, this is no easy task.

5.2. Example: A Complete Policy with All Customers Joining

THEOREM 9. Consider an M/M/s/N ($N \leq \infty$) system with balking option and priority option, and a policy defined by the decisions: $\Delta(\nu,R,h) = 1$ for $\nu = -s+1,\ldots,N$, and $b = b(\nu,R,h) \in B$ for $\nu = 1,\ldots,N$. Assume that $R - b(\nu,R,h) - h/\mu > 0$ for $\nu \leq 0$, and all $(R,h) \in \Omega$.

The policy is FIFO stable if and only if $N < \infty$, (i), (ii), (iii) of Theorem 5 are satisfied, and furthermore, (iv) $R - b_o - h[1/\mu + N/(s\mu)] \underset{(=)}{\geq} 0$ for all $(R,h) \in \Omega$, with $>$ only for strict stability.

The policy is LIFO-stable if and only if $N < \infty$, (i), (ii), (iii) of Theorem 7 are satisfied, and furthermore, (iv) $R - b_\nu - h[1/\mu + [1/(s\mu)] \cdot \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i] \underset{(=)}{\geq} 0$ for all $(R,h) \in \Omega$ and all $\nu = 1,\ldots,N$, with $>$ only for strict stability.

PROOF. As usual we shall give a proof only for the case of strong stability. Note, however, exactly the same words apply to the case of weak stability.

By assumption we have $R - b(\nu,R,h) - h/\mu > 0$ for $\nu \leq 0$, so stability for all ν-customers with $\nu \leq 0$ is automatically ensured. Hence we shall concern ourselves only with the question of stability for ν-customers with $\nu \geq 1$.

Suppose $N < \infty$. Stability means two things, namely stability

against payments other than the policy payment, and stability against balking.

FIFO-stability with respect to other payments obviously is attained if and only if conditions (i), (ii), (iii) of Theorem 5 (assuming forced joining) are satisfied. Given (i), stability with respect to balking will be achieved if and only if $R-b_o-h[1/\mu + \nu/(s\mu)] \geq 0$ for all $(R,h) \in \Omega$, $\nu = 1,\ldots, N$, that is if $R-b_o-h[1/\mu + N/(s\mu)] \geq 0$ for all $(R,h) \in \Omega$. Thus the stated conditions are necessary and sufficient for FIFO-stability.

LIFO-stability with respect to other payments obviously is attained if and only if conditions (i), (ii), (iii) of Theorem 7 (assuming forced joining) are satisfied. Given (i) and (ii), stability with respect to balking will be achieved if and only if $R-b_\nu-h[1/\mu + [1/(s\mu)] \sum_{i=0}^{N-\nu} (\lambda/(s\mu))^i] \geq 0$ for all $(R,h) \in \Omega$ and $\nu = 1,\ldots,N$. Thus the stated conditions are necessary and sufficient for LIFO-stability.

Finally, $N = \infty$ precludes FIFO-stability as well as LIFO-stability. FIFO-stability of the policy clearly is out of the question since, for each h, the expected queueing cost increases without a bound as ν increases, i.e. $\lim_{\nu\to\infty} h\nu/(s\mu) = \infty$. For LIFO-stability the conditions of Theorem 8 must be satisfied. Hence $b_{\nu+1} - b_\nu > c > 0$ for all ν, so $\lim_{\nu\to\infty} b_\nu = \infty$. In both cases, for each (R,h) there is a greatest ν at which it is profitable to join. Thus, with a balking option, a stable policy as specified cannot exist. □

6. Appendix

Here we shall discuss various natural choices of definitions of stability and reduced action sets. Relations between the set of stable

points and the reduced action space are derived. Our main objective is to explore how the reduced action space depends upon reduction scheme and dominance criterion. Among other things it is shown that under our model's strong dominance criterion (with preference rule a factor) and corresponding strong stability concept, the reduced action space is unique, that is the same for all admissible reduction schemes, and all strongly stable points lie in the reduced action space.

6.1. Definitions

Let A_ℓ^1 (ℓ=1,n) denote the complete action sets. We refer to an action combination $(a_1,\ldots,a_n)$, $a_\ell \in A_\ell^1$ (ℓ=1,n), as a point in the (complete) action space

$$S = \bigtimes A_\ell^1 = \{(a_1,\ldots,a_n): \quad a_\ell \in A_\ell^1, \ \ell=1,n\}.$$

Let S_k denote the set of all possible combinations of actions by players other than k, that is

$$S_k = \bigtimes_{\ell\neq k} A_\ell^1 = \{(a_1,\ldots,a_{k-1},a_{k+1},\ldots,a_n): \quad a_\ell \in A_\ell^1, \ \ell=1,\ldots,k-1,k+1,\ldots,n\}. \qquad (k=1,n)$$

Let $v_k(a_k;t)$ be player k's payoff when he chooses $a_k \in A_k^1$ while others choose $t \in S_k$. Let there also be given a preference rule which, for each k, provides a strict, simple ordering of all actions in A_k^1. Assume $T_k \subset S_k$, $T_k \neq \emptyset$. Comparing $a_k' \in A_k^1$ and $a_k'' \in A_k^1$, $a_k' \neq a_k''$, we say:

a_k' <u>strictly dominates</u> a_k'' on T_k if $v_k(a_k';t) > v_k(a_k'';t)$ for all $t \in T_k$.

a_k' <u>strongly dominates</u> a_k'' on T_k if $v_k(a_k';t) \geq v_k(a_k'';t)$ for all $t \in T_k$ and, in case $v_k(a_k';t) = v_k(a_k'';t)$ for any $t \in T_k$, then a_k' is preferred to a_k''.

a'_k normally dominates a''_k on T_k if $v_k(a'_k;t) \geq v_k(a''_k;t)$ for all $t \in T_k$ and $v_k(a'_k;t) > v_k(a''_k;t)$ for some $t \in T$.

We shall be concerned only with cases where player k considers his payoffs on the subspace $\bigtimes A_\ell$, where $A_\ell \subset A^1_\ell$. Thus $T_k = \bigtimes_{\ell \neq k} A_\ell$, $a'_k \in A_k$, $a''_k \in A_k$. For convenience we may then say "a'_k dominates a''_k on $\{A_\ell\}$."

Obviously, strict dominance implies strong dominance as well as normal dominance. Strong dominance is our dominance concept. Normal dominance is the dominance concept usually encountered in the literature.

Significantly, strict and strong dominance are preserved under set reduction. That is, if $T^1_k \subset T^2_k \subset S_k$ and a'_k dominates a''_k on T^2_k, then a'_k dominates a''_k also on T^1_k. In contrast normal dominance is not, in general, preserved under set reduction. This fact, easy to verify, has some undesirable consequences as we shall see.

From these three dominance concepts derive two stability concepts to be defined next. Let $a^o_\ell \in A^1_\ell$ $(\ell=1,n)$, and let (a^o_ℓ) denote the vector $(a^o_1,\ldots,a^o_n)$. We say:

(a^o_ℓ) is strictly stable if $v_k(a^o_1,\ldots,a^o_k,\ldots,a^o_n) > v_k(a^o_1,\ldots,a_k,\ldots,a^o_n)$ for all $a_k \in A^1_k - a^o_k$, all k.

(a^o_ℓ) is strongly stable if $v_k(a^o_1,\ldots,a^o_k,\ldots,a^o_n) > v_k(a^o_1,\ldots,a_k,\ldots,a^o_n)$, or, $v_k(a^o_1,\ldots,a^o_k,\ldots,a^o_n) = v_k(a^o_1,\ldots,a_k,\ldots,a^o_n)$ but a^o_k is preferred to a_k, for all $a_k \in A^1_k - a^o_k$, all k.

It is easy to see that strict stability means that for each k $(k=1,n)$ a^o_k strictly (and normally) dominates all $a_k \in A^1_k - a^o_k$ on $t^o_k = (a^o_1,\ldots,a^o_{k-1},a^o_{k+1},\ldots,a^o_n)$. Similarly, strong stability means that for each k $(k=1,n)$ a^o_k strongly dominates all $a_k \in A^1_k - a^o_k$ on t^o_k. In view of the first relationship strict stability might also be termed normal

stability. This however would be a misnomer, since strict stability, just like strong stability, is not widely used. The commonly used stability concept will here be called weak stability. It is defined as follows:

(a_ℓ^o) is <u>weakly stable</u> if $v_k(a_1^o,\ldots,a_k^o,\ldots,a_n^o) \geq v_k(a_1^o,\ldots,a_k,\ldots,a_n^o)$ for all $a_k \in A_k^1 - a_k^o$, all k.

Thus weak stability is the condition that for no k is a_k^o strictly (or normally) dominated by any $a_k \in A_k^1 - a_k^o$ on t_k^o.

It is clear from these definitions that strict stability implies strong stability which in turn implies weak stability. Our analysis employs the concept of strong stability extensively. However, results involving the other stability concepts are also presented.

6.2. Stable Points and Reduced Action Space

We begin by introducing some new terminology and notation. First, a point (a_ℓ^o) which satisfies the condition for stability we call a stable point, an equilibrium point, or, in the context of decision-making, a stable solution.

Let P(1), P(2), P(3) denote the sets of strictly, strongly and weakly stable points, respectively. From the definitions, $P(1) \subset P(2) \subset P(3) \subset S$.

Also, we choose to indicate by an argument of the reduced action set which dominance criterion has been employed. Let 1, 2, 3 indicate the use of strict, strong and normal dominance, respectively. Thus, $A_k^*(2)$ is the reduced action set of player k, obtained under strong dominance and an unspecified admissible reduction scheme; and $\times A_\ell^*(2)$ is the reduced action space.

We are interested in the relation between P(1), P(2) or P(3) on one

hand, and $\bigtimes A_\ell^*(1)$, $\bigtimes A_\ell^*(2)$, or $\bigtimes A_\ell^*(3)$ on the other. The results are:

$$P(1) \subset \bigtimes A_\ell^*(3) \tag{17}$$

$$P(1) \subset P(2) \subset \bigtimes A_\ell^*(2) \tag{18}$$

$$P(1) \subset P(3) \subset P(3) \subset \bigtimes A_\ell^*(1) \tag{19}$$

Proofs are easy. Consider for instance the case of the strong dominance criterion being used in the reduction process, resulting in $\bigtimes A_\ell^*(2)$. Suppose (a_ℓ^o) is strongly stable, i.e. $(a_\ell^o) \in P(2)$. It is immediately clear that none of the actions a_ℓ^o, $\ell=1,\ldots,n$, may ever be deleted under a strong dominance criterion, at any stage of the reduction process. Hence $(a_\ell^o) \in \bigtimes A_\ell^*(2)$, so that $P(2) \subset \bigtimes A_\ell^*(2)$. As $P(1) \subset P(2)$, this proves (18). Proofs of (17) and (19) are similar and will therefore be omitted.

Note, (17) and (18) cannot be strengthened. It is not true in general that P(2) or P(3) are contained in a "weakly" reduced action space, nor that P(3) is contained in a "strongly" reduced action space, see Section 6.4.

Suppose the reduced action space consists of a single point (a_ℓ^*). Is it stable? Again, the answer is simple and proofs are straightforward.

Reduction by strict dominance $\Rightarrow$ (a_ℓ^*) is strictly stable. (20)

Reduction by strong dominance $\Rightarrow$ (a_ℓ^*) is strongly stable. (21)

Reduction by normal dominance $\Rightarrow$ (a_ℓ^*) is weakly stable. (22)

Consider here the case of strong dominance criterion. By virtue of the fact that (a_ℓ^*) is the last remaining point, evidently $v_k(a_k^*;t_k^*) > v_k(a_k;t_k^*)$ for all $a_k \in A_k^1 - a_k^*$, all k, with $t_k^* = (a_1^*,\ldots,a_{k-1}^*,a_{k+1}^*,\ldots,a_n^*)$. But this

condition is precisely the definition of strict stability. This proves (21). Other proofs are similar, and will be omitted.

6.3. Effects of Reduction Scheme and Dominance Criterion

We now address the important question of how the reduced action sets depend on reduction scheme and on dominance criterion used in the reduction process. We shall prove that the reduced action space $\times A_\ell^*$ is unique in the class of all admissible reduction schemes under either strict or strong dominance. The proof has been generalized in a way that allows us to compare the reduced action sets for different reduction schemes and dominance criteria. We thereby learn that the (unique) $\times A_\ell^*$ obtained under strict dominance is at least as large as the (unique) $\times A_\ell^*$ obtained under strong dominance and at least as large as all of the $\times A_\ell^*$ obtained under normal dominance in the class of admissible reduction schemes.

Let the complete action sets $\{A_\ell^1\}$ be arbitrary. Further, let $A_k^{ij}(d)$ denote the partially reduced action set of player k at stage i (prior to deletion of dominated actions) under reduction scheme j and dominance criterion d, and let $R_k^{ij}(d)$ denote the actions deleted by player k at stage i under scheme j and dominance criterion d (k=1,...,n; i=1,2, ...; $j \in J \equiv$ the class of admissible reduction schemes; d = 1 (strict), 2 (strong), 3 (normal)). $A_k^{1j}(d) \equiv A_k^1$. The corresponding reduced action sets are symbolized by $A_k^{*j}(d)$.

First, it will be shown that

$$a_k \in \bigcup_{i=1}^{\infty} R_k^{ij_1}(1) \Rightarrow a_k \in \bigcup_{i=1}^{\infty} R_k^{ij_2}(d) \tag{23}$$

for k=1,...,n; all $a_k \in A_k^1$; $j_1 \in J$, $j_2 \in J$; d = 1,2,3. Thus we claim that

if a_k is dominated at any stage of reduction process 1, defined by scheme j_1 and strict dominance, then it must also be dominated at some stage of reduction process 2, defined by scheme j_2 and any of the three dominance criteria, labeled d.

First consider the simplest case, $a_k \in R_k^{1j_1}(1)$. Then $a_k \in A_k^1$ is strictly dominated on $\{A_\ell^{1j_1}(1)\} \equiv \{A_\ell^1\}$. In process 2 let player k have his first turn at stage i_1, $0 < i_1 \leq L$, where L is the upper limit associated with scheme j_2. At stage i_1 the partially reduced action sets are $\{A_\ell^{i_1 j_2}(d)\} = \{A_\ell^1 - \bigcup_{i=0}^{i_1-1} R_\ell^{ij_2}(d)\}$ with $R_\ell^{0j_2}(d) \equiv \emptyset$. Now, by definition of i_1, $A_k^{i_1 j_2}(d) = A_k^1$, and since a strict dominance relation between two actions in A_k^1 is not affected by the reduction from $\{A_\ell^{1j_1}(1)\} \equiv \{A_\ell^1\}$ to $\{A_\ell^{i_1 j_2}(d)\}$ we conclude that a_k is strictly dominated, hence d-dominated, on $\{A_\ell^{i_1 j_2}(d)\}$. It follows that

$$a_k \in R_k^{1j_1}(1) \Rightarrow a_k \in R_k^{i_1 j_2}(d) \subset \bigcup_{i=1}^{\infty} R_k^{ij_2}(d) \tag{24}$$

Next consider the case $a_k \in R_k^{2j_1}(1)$. Then $a_k \in A_k^{2j_1}(1)$ is strictly dominated on $\{A_\ell^{2j_1}(1)\}$. Assume a_k is strictly dominated by $a_k' \in A_k^{2j_1}(1)$. Suppose that in process 2 player k's first turn following stage L is at stage i_2, $L < i_2 \leq 2L$. If at this stage a_k has already been deleted, then there is no problem, so assume $a_k \in A_k^{i_2 j_2}(d)$. By (24) all actions deleted at stage 1 of process 1 will be deleted no later than at stage L of process 2 and hence will have been deleted at stage $i_2 > L$. That is, $A_\ell^1 - \bigcup_{i=1}^{i_2-1} R_\ell^{ij_2}(d) \subset A_\ell^1 - R_\ell^{1j_1}(1)$, or $A_\ell^{i_2 j_2}(d) \subset A_\ell^{2j_1}(1)$, for $\ell = 1,\ldots,n$. Thus, at stage i_2 of process 2 player k faces an action space $\times A_\ell^{i_2 j_2}(d)$ which is smaller than or equal to the one considered at stage 2 of

process 1, namely $\times A_{\ell}^{2j_1}(1)$. By assumption, $a_k \varepsilon A_k^{i_2 j_2}(d)$. If also $a_k' \varepsilon A_k^{i_2 j_2}(d)$ then a_k must be strictly dominated by a_k' on $\{A_{\ell}^{i_2 j_2}(d)\}$ since strict dominance relations carry over to a smaller action space. If, however, $a_k' \notin A_k^{i_2 j_2}(d)$ then there is an $a_k'' \varepsilon A_k^{i_2 j_2}(d)$ which strictly dominates a_k on $\{A_{\ell}^{i_2 j_2}(d)\}$ because otherwise a_k' could not have been deleted. In either case, a_k is strictly dominated, hence d-dominated, on $\{A_{\ell}^{i_2 j_2}(d)\}$. Thus, always, $a_k \varepsilon R_k^{2j_1}(1) \Rightarrow a_k \varepsilon \overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_2}(d)$.

In exactly the same way one proves that $a_k \varepsilon R_k^{rj_1}(1) \Rightarrow a_k \varepsilon \overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_2}(2)$ for r=3 and by induction the relation is shown to hold for all r = 1,2,.... Hence (23) is true.

Now, (23) implies $\overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_1}(1) \subset \overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_2}(d)$. Since $A_k^{*j_1}(1) = A_k^1 - \overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_1}(1)$ and $A_k^{*j_2}(d) = A_k^1 - \overset{\infty}{\underset{i=1}{\cup}} R_k^{ij_2}(d)$ we deduce

$$A_k^{*j_2}(d) \subset A_k^{*j_1}(1) \qquad (k=1,\ldots,n;\ \text{all } j_1,j_2 \varepsilon J;\ d=1,2,3). \tag{25}$$

Also,

$$A_k^{*j_2}(2) \subset A_k^{*j_1}(2) \qquad (k=1,\ldots,n;\ \text{all } j_1,j_2 \varepsilon J). \tag{26}$$

Relation (26) is proved by the same method by which (25) was obtained The key factor in the proof of (26) is that strong dominance is preserved under reduction of the action space.

The most important conclusions to be drawn from (25) and (26) are

$$A_k^{*j_1}(1) = A_k^{*j_2}(1) \equiv A_k^*(1) \qquad (k=1,\ldots,n;\ \text{all } j_1,j_2 \varepsilon J), \tag{27}$$

$$A_k^{*j_1}(2) = A_k^{*j_2}(2) \equiv A_k^*(2) \qquad (k=1,\ldots,n;\ \text{all } j_1,j_2 \varepsilon J). \tag{28}$$

These equations state that when the dominance criterion used in reduction is either strict or strong dominance then the reduced action sets are

unique with respect to the class of all admissible reduction schemes. Briefly stated, $\{A_\ell^*(1)\}$ and $\{A_\ell^*(2)\}$ are unique.

By using (27) and (28) in (25) we find

$$A_k^*(2) \subset A_k^*(1) \qquad (k=1,\ldots,n), \tag{29}$$

$$A_k^{*j}(3) \subset A_k^*(1) \qquad (k=1,\ldots,n;\ \text{all } j\varepsilon J). \tag{30}$$

Neither $A_k^{*j}(3) \subset A_k^*(2)$ nor $A_k^*(2) \subset A_k^{*j}(3)$ hold in general, see Section 6.4.

6.4. Counterexample

For a simple example which, among other things, provides proof of nonuniqueness of $\{A_\ell^*\}$ under normal dominance, consider a two-person game in which each player has two alternatives, actions 1 and 2. Let the payoffs be

$$v_1(1,1) = -v_2(1,1) = 1 \qquad v_1(1,2) = -v_2(1,2) = 0$$
$$v_1(2,1) = -v_2(2,1) = 2 \qquad v_1(2,2) = -v_2(2,2) = 2$$

Consider two different preference rules. By rule 1 action 1 is preferred to action 2 (both players). By rule 2 action 2 is preferred to action 1.

The stable points are easily located as follows:

Strictly stable points: $P(1) = \emptyset$.

Strongly stable points under rule 1: $P'(2) = \{(2,1)\}$.

Strongly stable points under rule 2: $P''(2) = \{(2,2)\}$.

Weakly stable points: $P(3) = \{(2,1),(2,2)\}$.

For the case of normal dominance we shall discuss three types of reduction schemes. Under scheme 1, described I, II, ..., player 1 reduces

at stage 1, player 2 reduces at stage 2. Under scheme 2, described II, I, ..., player 2 reduces at stage 1, player 1 reduces at stage 2. Finally, under scheme 3, described (I, II),..., both players reduce at stage 1.

In Table III-1 we list the reduced action sets under various assumptions as to preference rule, dominance criterion applied in the reduction process, and reduction scheme.

Table III-1. Reduced Action Sets for Various Reduction Processes

dominance concept	preference rule	reduction scheme	A_1^*	A_2^*
strict (1)	-	any $j \in J$	{2}	{1,2}
strong (2)	rule 1	any $j \in J$	{2}	{1}
	rule 2	any $j \in J$	{2}	{2}
normal (3)	-	scheme 1	{2}	{1,2}
		scheme 2	{2}	{2}
		scheme 3	{2}	{2}

First, compare reduced action sets. The table shows that, in general, $\{A_\ell^*(3)\}$ is not unique. It also shows that, in general, neither $A_\ell^*(2) \subset A_\ell^*(3)$ nor $A_\ell^*(3) \subset A_\ell^*(2)$ holds true. This can be ascribed to the fact that strong dominance does not, in general, imply normal dominance, nor does normal dominance imply strong dominance.

Second relate stable points and action sets. Our example disproves the statements $P(3) \subset \bigtimes A_\ell^*(2)$ and $P(3) \subset \bigtimes A_\ell^*(3)$. That is, weakly stable points do not necessarily lie in a reduced action space generated in a process based on strong or normal dominance.

CHAPTER IV

AN M/M/1 QUEUE WITH TWO USERS CHOOSING ARRIVAL RATES

Our subject is decision-making in an M/M/1 queue with two users. The queue is fed by two independent Poisson streams. The arrival rate λ_i for stream i (i=1,2) is under the exclusive control of user i with $0 \leq \lambda_i < \infty$. Thus, the interarrival times for stream i are independent, exponentially distributed variables with distribution function $F_i(t) = 1-e^{-\lambda_i t}$. The sum of two independent Poisson streams is a Poisson stream with arrival rate equal to the sum of the two arrival rates. Hence the interarrival times in the combined stream are independent, exponentially distributed variables with distribution function $F(t) = 1-e^{-(\lambda_1+\lambda_2)t}$. All arrivals are served on a first-come-first-served basis. Service times are independent, exponentially distributed variables with distribution function $H(t) = 1-e^{-\mu t}$, corresponding to service rate μ and mean service time μ^{-1}.

A user receives a reward from the service of "his" customers. Let $R_i(\lambda_i)$ be user i's reward per unit time under arrival rate λ_i. In order to avoid trivialities assume $R_i(\lambda_i) > R_i(0)$ for some $\lambda_i > 0$. Also, assume $R_i(\ )$ is continuous and twice differentiable on $[0,\infty)$ with continuous derivatives, and that there exists an upper bound on the first derivative. Thus $M_i' \equiv \sup_{0 \leq \lambda_i < \infty} R_i'(\lambda_i) < \infty$ (i=1,2). As $R_i(\lambda_i) > R_i(0)$ for some $\lambda_i > 0$ we have $M_i' > 0$ (i=1,2). Constraints on the second derivative will be introduced later.

On the other hand, a user will incur a waiting loss when $\lambda_i > 0$. For user i this loss equals $h_i > 0$ per unit time a customer spends in the system. The corresponding long run average loss per unit time is $h_i\lambda_i w(\lambda_1+\lambda_2;\mu)$ where $w(\lambda;\mu)$ is the expected waiting time (in system) of a customer when λ $(=\lambda_1+\lambda_2>0)$ is the overall arrival rate and μ is the service rate.

The objective of each user is to maximize the difference between reward and waiting loss. Often we shall refer to this difference as the profit.

Standard queueing theory, see for example Kleinrock [1975, p. 98], tells us that in the M/M/1 queue the waiting time in system has the mean value ∞ if $\lambda \geq \mu$, and $1/(\mu-\lambda)$ if $0 < \lambda < \mu$ (actually, the waiting time is here exponentially distributed). Hence, the objective function of user i (i=1,2) is

$$\pi_i(\lambda_i;\mu,\lambda_j) = \begin{cases} R_i(0) & (\lambda_i=0), \\ R_i(\lambda_i)-h_i\lambda_i/(\mu-\lambda_1-\lambda_2) & (\lambda_i>0,\ \lambda_1+\lambda_2<\mu), \\ -\infty & (\lambda_i>0,\ \lambda_1+\lambda_2\geq\mu). \end{cases} \qquad (1)$$

Here as in the sequel we use the convention of letting j designate the other user. Thus (i,j) equals (1,2) or (2,1). By (1), $\pi_i(\lambda_i;\mu,\lambda_j)$ depends on μ and λ_j only through their difference, $\mu-\lambda_j$. Therefore, as far as user i is concerned the effect of the presence of user j is exactly the same as a decrease in queue capacity, from μ to $\mu-\lambda_j$.

In (1) it was understood that $\mu > 0$. In order to facilitate the statement of some results the definition of π is extended to include

nonpositive μ as follows.

$$\pi_i(\lambda_i;\mu,\lambda_j) \equiv \begin{cases} R_i(0) & (\lambda_i=0,\ \mu\leq 0), \\ -\infty & (\lambda_i>0,\ \mu\leq 0). \end{cases} \tag{2}$$

Then, by (1), we have the simple relation

$$\pi_i(\lambda_i;\mu,\lambda_j) = \pi_i(\lambda_i;\mu+c,\lambda_j+c) \qquad (c\geq -\lambda_j). \tag{3}$$

In particular,

$$\pi_i(\lambda_i;\mu,\lambda_j) = \pi_i(\lambda_i;\mu-\lambda_j,0). \tag{4}$$

For simplicity define $\phi_i(\lambda_i;\mu) \equiv \pi_i(\lambda_i;\mu,0)$. Then (4) can be written

$$\pi_i(\lambda_i;\mu,\lambda_j) = \phi_i(\lambda_i;\mu-\lambda_j). \tag{5}$$

Equation (5) explains the importance of an analysis of the one-user model as a preliminary step to the analysis of the two-user model. For $\mu > 0$, $\phi_i(\lambda_i;\mu)$ is precisely the profit of user i in a one-user model, where user j's arrival rate is identically zero. Specifically, setting $\lambda_j = 0$ in (1), we have for $\mu > 0$

$$\phi_i(\lambda_i;\mu) = \begin{cases} R_i(\lambda_i)-h_i\lambda_i/(\mu-\lambda_i) & (\lambda_i<\mu), \\ -\infty & (\lambda_i\geq\mu). \end{cases} \tag{6}$$

The one-user model is analyzed in Section 1. Our main objective is the determination of the optimal arrival rate $\overline{\lambda}$ and the associated profit $\overline{\phi} = \phi(\overline{\lambda};\mu)$. We also explore $\overline{\lambda}$'s and $\overline{\phi}$'s dependence on μ and h. The

special case, $R(\lambda) = \lambda r$ is used for illustration.

Sections 2, 3, and 4 discuss the two-user model under different behavioral assumptions.

In Section 2 we begin by defining an equilibrium point $(\hat{\lambda}_1,\hat{\lambda}_2)$ of the decision variables. It is shown that under quite general conditions an equilibrium point will exist and that it is unique and dynamically stable (globally). If so, $(\hat{\lambda}_1,\hat{\lambda}_2)$ is the solution in a reciprocal rate adjustment model we term follower-follower. $(\hat{\lambda}_1,\hat{\lambda}_2)$ and associated profits $(\hat{\pi}_1,\hat{\pi}_2)$ are compared with $(\bar{\lambda}_1,\bar{\lambda}_2)$ and $(\bar{\phi}_1,\bar{\phi}_2)$ in the corresponding one-user models. Furthermore, we investigate how $(\hat{\lambda}_1,\hat{\lambda}_2)$ and $(\hat{\pi}_1,\hat{\pi}_2)$ depend on μ, h_1 and h_2. The special case, $R_i(\lambda_i) = r_i\lambda_i$ $(i=1,2)$ is used for illustration.

Section 3 contains an analysis of the leader-follower decision model in which the follower automatically adjusts his arrival rate to that of the leader who sets his rate, once and for all, so that his profit is maximized taking into account the follower's response. Calling the optimizing arrival rates $(\lambda^*_{i1},\lambda^*_{j2})$ and associated profits (π^*_{i1},π^*_{j2}) for leader (subscript 1) and follower (subscript 2), respectively, it is shown that $\lambda^*_{i2} \leq \hat{\lambda}_i \leq \lambda^*_{i1}$ and $\pi^*_{i2} \leq \hat{\pi}_i \leq \pi^*_{i1}$ for $i = 1,2$. Again, we illustrate with the case $R_i(\lambda_i) = r_i\lambda_i$ $(i=1,2)$.

In Section 4 we discuss the cooperative model in which the combined profits $\psi = \pi_1 + \pi_2$ are maximized by choice of λ_1 and λ_2. The solution of this model is compared to the solution of the follower-follower model. Here also, we use the case $R_i(\lambda_i) = r_i\lambda_i$ $(i=1,2)$ for illustration.

It appears that our two-user model has never been the subject of analysis in the queueing literature. Thus no literature references

concerning the model can be given. However, there are other models in which customers or users make decisions affecting arrival rates, with no balking allowed.

Littlechild [1974] examines decision-making by individual customers in an M/M/1 queue. On the basis of information on the total arrival rate λ, each customer decides whether to arrive or not (balking is not allowed). The expected gain from service is $R-h/(\mu-\lambda)$, where h is constant, but R is a variable with continuous distribution function. It is shown that a unique equilibrium arrival rate λ_o exists. The essential difference between ours and Littlechild's model is that we have only two users whereas he has infinitely many (= customers). The effect is a radically different model. For one thing, no customer sets an arrival rate and the question of behavioral assumption does not arise.

Balachandran and Schaefer [1976[analyze Littlechild's model with the difference that customers fall into k classes with constant (R_i,h_i) $(i=1,k)$ instead of being characterized by constant h and continuously distributed R. Another paper by the same authors [unpublished] deals with the M/G/1 queue.

Closely related models appear in the economics literature under the names of duopsony (two buyers) and duopoly (two sellers). For a discussion of these models, see Henderson and Quandt [1971] and Intriligator [1971]. Our assumptions are stronger than those of the general theory. Hence we have many results that do not follow from the available literature on duopsony and duopoly.

1. One User

In this section there is no need for a subscript identifying the user so we leave it out and write the objective function, (6), as

$$\phi(\lambda;\mu) = \begin{cases} R(\lambda)-h\lambda/(\mu-\lambda) & (\lambda<\mu), \\ -\infty & (\lambda\geq\mu). \end{cases} \tag{7}$$

Evidently, no $\lambda\geq\mu$ can ever maximize ϕ so the user need consider only values of λ on the interval $[0,\mu)$. By our assumptions about R, ϕ will be continuous and twice differentiable on $[0,\mu)$. $\phi(0;\mu) = R(0) < \infty$. Also, as $R'(\lambda) < M'$ it follows that $\phi(\lambda;\mu) < R(0)+\lambda(M'- h/(\mu-\lambda))$, so $\lim_{\lambda\to\mu} \phi(\lambda;\mu) = -\infty$. Hence there exists a λ_o, $0 < \lambda_o < \mu$, such that $\phi(\lambda;\mu) < \phi(0;\mu)$ for $\lambda \geq \lambda_o$. We deduce that ϕ has a global maximum $\overline{\phi}$ for at least one $\lambda \in [0,\lambda_o]$ and hence for at least one $\lambda \in [0,\mu)$. We assume that the user chooses the minimal λ satisfying $\phi(\lambda;\mu) = \overline{\phi}$ and call it $\overline{\lambda}$.

Differentiating ϕ twice with respect to λ we derive

$$\phi'(\lambda;\mu) \equiv d\phi(\lambda;\mu)/d\lambda = R'(\lambda)-h\mu/(\mu-\lambda)^2 \qquad (\lambda<\mu), \tag{8}$$

$$\phi''(\lambda;\mu) \equiv d^2\phi(\lambda;\mu)/d\lambda^2 = R''(\lambda)-2h\mu/(\mu-\lambda)^3 \qquad (\lambda<\mu). \tag{9}$$

As is well known, $\phi'(\lambda;\mu) = 0$ is a necessary condition, and $\phi'(\lambda;\mu) = 0$, $\phi''(\lambda;\mu) < 0$ are sufficient conditions for ϕ to possess a local maximum at $\lambda \in (0,\mu)$. Later we shall impose a condition on $R''(\,)$ under which $\phi'(\overline{\lambda};\mu)=0$ is sufficient for a unique global maximum at $\overline{\lambda}$, i.e. $\phi(\overline{\lambda};\mu) > \phi(\lambda;\mu)$ for all $\lambda \neq \overline{\lambda}$.

1.1. General Results

We shall begin by presenting a useful upper bound on $\overline{\lambda}$. By (8) and $R'(\lambda) \leq M'$,

$$\phi'(\lambda;\mu) \leq M'-h\mu/(\mu-\lambda)^2 \qquad (\lambda<\mu).$$

Solving $M'-h\mu/(\mu-\lambda)^2 = 0$ for λ results in $\lambda = \mu - \sqrt{\mu h/M'}$. Hence

$$\phi'(\lambda;\mu) < 0 \qquad (\mu-\sqrt{\mu h/M'} < \lambda < \mu).$$

Since $\phi'(\lambda;\mu) < 0$ precludes optimality, except at $\lambda = 0$, we conclude that

$$\overline{\lambda} \leq \max(0,\mu-\sqrt{\mu h/M'}). \tag{10}$$

At this point few constraints have been placed on R and not much of value can be said on the determination of $\overline{\lambda}$ and $\overline{\phi}$. It is of interest, however, in what way $\overline{\lambda}$ and $\overline{\phi}$ respond to change in the parameters μ and h. Lemma 1 and Lemma 2 below state our findings.

LEMMA 1. For given R and h, there exists a μ^*, $0 < \mu^* < \infty$, such that $\overline{\lambda} = 0$ if and only if $\mu \leq \mu^*$. For $\mu \in [\mu^*,\infty)$, $\overline{\phi}$ is continuous and strictly increasing in μ; and $\overline{\lambda}$ is strictly increasing in μ, but is not necessarily continuous in μ.

LEMMA 2. For given R and μ, there exists an h^*, $0 \leq h^* < \infty$, such that $\overline{\lambda} = 0$ if and only if $h \geq h^*$. For $h \in (0,h^*]$, $\overline{\phi}$ is continuous and strictly decreasing in h; and $\overline{\lambda}$ is strictly decreasing in h, but is not necessarily continuous in h.

Proofs are presented in Section 5. We hint only that a key to the proof is that the waiting loss function $z = h\lambda/(\mu-\lambda)$ for $0 < \lambda < \mu$ has the properties

$$dz/d\mu = - h\lambda/(\mu-\lambda)^2 < 0, \tag{11}$$

$$d^2z/d\mu d\lambda = - h(\mu+\lambda)/(\mu-\lambda)^3 < 0, \tag{12}$$

$$dz/dh = \lambda/(\mu-\lambda) > 0, \tag{13}$$

$$d^2z/dhd\lambda = \mu/(\mu-\lambda)^2 > 0. \tag{14}$$

Note, by (10), $\overline{\lambda} = 0$ if $\mu-\sqrt{\mu h/M'} \leq 0$. Thus $\overline{\lambda} = 0$ if $\mu \leq h/M'$, or, equivalently, if $h \geq \mu M'$. Hence, by Lemma 1,

$$\mu^* \geq h/M', \tag{15}$$

and, by Lemma 2,

$$h^* \leq \mu M'. \tag{16}$$

1.2. The Extra Condition on R

Henceforth we shall impose a condition on R which will ensure that ϕ has a unique global maximum and, connected therewith, that $\overline{\lambda}$ is continuous in μ and h. Condition 1 is that condition. Other conditions we may use, namely Conditions 2 and 3, are stronger conditions which imply Condition 1. For easy reference we state all of these conditions here.

CONDITION 1. $R''(\lambda) < 2h\mu/(\mu-\lambda)^3$ (all $\lambda<\mu$). (17)

CONDITION 2. $R''(\lambda) < h/(\mu-\lambda)^2$ (all $\lambda<\mu$). (18)

CONDITION 3. $R''(\lambda) \leq 0$ (all $\lambda\geq 0$). (19)

By (9), Condition 1 means nothing but $\phi''(\lambda;\mu) < 0$ for all $\lambda < \mu$. That is, the condition imposes strict concavity on ϕ on $[0,\mu)$. We have already found that ϕ has a global maximum on $[0,\mu)$. By a fundamental property of strictly concave functions, Condition 1 guarantees that ϕ has a unique global maximum for some $\overline{\lambda} \in [0,\mu)$. If $\phi'(0;\mu) \leq 0$, then $\overline{\lambda} = 0$, and if $\phi'(0;\mu) > 0$, then $\overline{\lambda} > 0$ and it satisfies $\phi'(\overline{\lambda};\mu) = 0$.

Note, Condition 3 ⇒ Condition 2 ⇒ Condition 1 ((19)⇒(18)⇒(17)), as $0 < h/(\mu-\lambda)^2 < 2h\mu/(\mu-\lambda)^3$ for all $\lambda < \mu$. Obviously, the satisfaction of Condition 3 does not depend on μ and h. Lemma 3 tells us how the satisfaction of Conditions 1 and 2 depends on the parameter μ. Lemma 4 tells us how the satisfaction of the conditions depends on h.

LEMMA 3. Let R and h be given. If $R''(\lambda) > 0$ for some $\lambda \geq 0$, then there exist $\mu^{(1)}$ and $\mu^{(2)}$, $0 < \mu^{(2)} < \mu^{(1)} < \infty$ such that $\mu < \mu^{(1)} \Leftrightarrow$ (17) and $\mu < \mu^{(2)} \Leftrightarrow$ (18). If $R''(\lambda) \leq 0$ for all $\lambda \geq 0$ then (17) and (18) hold for all μ. Setting $\mu^{(1)} = \mu^{(2)} = \infty$ the above equivalence relations hold also in this case.

LEMMA 4. Let R and μ be given. If $R''(\lambda) > 0$ for some λ, $0 \leq \lambda < \mu$, then there exist $h^{(1)}$ and $h^{(2)}$, $0 < h^{(1)} < h^{(2)} < \infty$, such that $h > h^{(1)}$ $\Leftrightarrow$ (17) and $h > h^{(2)} \Leftrightarrow$ (18). If $R''(\lambda) \leq 0$ for all λ, $0 \leq \lambda < \mu$, then (17) and (18) hold for all h. Setting $h^{(1)} = h^{(2)} = 0$ the above equivalence relations hold also in this case.

Proofs of the lemmas are given in Section 5. We shall mention only that the proof depends in an obvious way on the fact that, for any $\lambda < \mu$, both $2h\mu/(\mu-\lambda)^3$ and $h/(\mu-\lambda)^2$ are strictly decreasing in μ, and strictly increasing in h. In the following, Lemma 3 is of crucial importance. The result we need is that if Condition 1 holds for a given μ, then it will hold for all smaller μ's, and similarly for Condition 2.

1.3. The Solution $\bar{\lambda}$

Condition 1 has been assumed to hold. Hence, as previously stated, ϕ will have a unique global maximum for a $\bar{\lambda} \in [0,\mu)$. If $\phi'(0;\mu) \leq 0$, that is, by (8), if $R'(0) \leq h/\mu$, then since ϕ is strictly concave we

have $\phi(0;\mu) > \phi(\lambda;\mu)$ for all $\lambda > 0$, and therefore $\bar{\lambda} = 0$. Defining

$$\mu^* \equiv h/R'(0), \tag{20}$$

we have

$$\bar{\lambda} = 0 \qquad (\mu \leq \mu^*). \tag{21}$$

If, on the other hand, $\phi'(0;\mu) > 0$, that is if $\mu > \mu^*$, then ϕ has an interior maximum at $\bar{\lambda}$ $(0<\bar{\lambda}<\mu)$, which is the solution of $\phi'(\lambda;\mu) = 0$. Thus, by (8),

$$R'(\bar{\lambda}) = h\mu/(\mu-\bar{\lambda})^2 \qquad (\mu>\mu^*). \tag{22}$$

The corresponding ϕ, namely $\bar{\phi}$, must be calculated by insertion of $\bar{\lambda}$ into (7): $\bar{\phi} = R(\bar{\lambda}) - h\bar{\lambda}/(\mu-\bar{\lambda})$.

1.4. The Dependence of $\bar{\lambda}$ on μ and h

As mentioned, Condition 1 implies continuity of $\bar{\lambda}$ in μ and h. Lemma 5 contains the formal statement of this result. Proof is found in Section 5.

LEMMA 5. For given R and h, $\bar{\lambda}$ is continuous in μ on $(0,\mu^{(1)})$. For given R and μ, $\bar{\lambda}$ is continuous in h on $(h^{(1)},\infty)$.

We want to find out how $\bar{\lambda}$ is affected by a small change $d\mu$ for a μ satisfying Condition 1, i.e. $\mu < \mu^{(1)}$. If $\mu < \mu^*$, then, of course, $d\bar{\lambda}/d\mu = 0$, by (21). If $\mu > \mu^*$, then (22) applies, and $d\bar{\lambda}/d\mu$ is derived by implicit differentiation of (22) with respect to $\bar{\lambda}$ and μ. The result is

$$[R''(\bar{\lambda}) - 2h\mu/(\mu-\bar{\lambda})^3]d\bar{\lambda} + [h(\mu+\bar{\lambda})/(\mu-\bar{\lambda})^3]d\mu = 0,$$

by which

$$\frac{d\bar{\lambda}}{d\mu} = \frac{h(\mu+\bar{\lambda})/(\mu-\bar{\lambda})^3}{2h\mu/(\mu-\bar{\lambda})^3 - R''(\bar{\lambda})} > 0 \qquad (\mu^* < \mu < \mu^{(1)}). \tag{23}$$

We conclude that $\bar{\lambda}$ is continuous and differentiable in μ on, at least, the interval $[\mu^*, \mu^{(1)}]$.

Similarly, defining $h^* \equiv R'(0)\mu$, we obtain

$$\frac{d\bar{\lambda}}{dh} = \frac{\mu/(\mu-\bar{\lambda})^2}{R''(\bar{\lambda}) - 2h\mu/(\mu-\bar{\lambda})^3} < 0 \qquad (h^{(1)} < h < h^*), \tag{24}$$

and we conclude that $\bar{\lambda}$ is continuous and differentiable in h on, at least, the interval $[h^{(1)}, h^*]$.

1.5. A Condition for $0 < d\bar{\lambda}/d\mu < 1$

As we have seen, Condition 1 is sufficient for a unique $\bar{\lambda}$ and $d\bar{\lambda}/d\mu \geq 0$. Is there a simple, stronger condition that, in addition, guarantees $d\bar{\lambda}/d\mu < 1$? Assume Condition 1 is met, that is, $0 < \mu < \mu^{(1)}$. If $\mu < \mu^*$, then $0 = d\bar{\lambda}/d\mu < 1$. If $\mu^* < \mu < \mu^{(1)}$, the requirement $d\bar{\lambda}/d\mu < 1$, according to (23), translates into

$$\frac{h(\mu+\bar{\lambda})/(\mu-\bar{\lambda})^3}{2h\mu/(\mu-\bar{\lambda})^3 - R''(\bar{\lambda})} < 1.$$

Observe, under Condition 1, the denominator of the left-hand side fraction is positive. The above inequality reduces to

$$R''(\bar{\lambda}) < h/(\mu-\bar{\lambda})^2.$$

In Lemma 6 we state this and related results that are easily shown to hold.

LEMMA 6. Suppose Condition 1 holds, and $\mu > \mu^*$. Then

(i) $d\bar{\lambda}/d\mu < 1$ if and only if $R''(\bar{\lambda}) < h/(\mu-\bar{\lambda})^2$,

(ii) $d\bar{\lambda}/d\mu = 1$ if and only if $R''(\bar{\lambda}) = h/(\mu-\bar{\lambda})^2$,

(iii) $d\bar{\lambda}/d\mu > 1$ if and only if $R''(\bar{\lambda}) > h/(\mu-\bar{\lambda})^2$.

We are now in the position to state and prove a theorem that is of great importance in our analysis of the two-user model.

THEOREM 1. If Condition 2 holds and $\mu > \mu^* = h/R'(0)$, then $0 < d\bar{\lambda}/d\mu < 1$.

PROOF. Let $\mu > \mu^*$ and let $\bar{\lambda}$ be the corresponding optimal arrival rate. Assume furthermore Condition 2 holds. Then, by previous remarks, also Condition 1 will hold, so Lemma 6 applies. By Condition 2, in particular, $R''(\bar{\lambda}) < h/(\mu-\bar{\lambda})^2$. By Lemma 6, part (i), it follows that $d\bar{\lambda}/d\mu < 1$. The other inequality, $d\bar{\lambda}/d\mu > 0$, results from (23). □

1.6. The Dependence of $\bar{\phi}$ on μ and h

We want to find out how $\bar{\phi}$ is affected by a small change $d\mu$ for a μ satisfying Condition 1, i.e. $\mu < \mu^{(1)}$. If $\mu < \mu^*$, then $\bar{\phi} = R(0)$, so $d\bar{\phi}/d\mu = 0$. If $\mu > \mu^*$, then $\bar{\lambda} > 0$ and

$$\bar{\phi} = R(\bar{\lambda}) - h\bar{\lambda}/(\mu-\bar{\lambda}).$$

Differentiating $\bar{\phi}$ with respect to μ we derive

$$\frac{d\bar{\phi}}{d\mu} = R'(\bar{\lambda})\frac{d\bar{\lambda}}{d\mu} - h(\mu\frac{d\bar{\lambda}}{d\mu}-\bar{\lambda})/(\mu-\bar{\lambda})^2.$$

Insertion of $R'(\bar{\lambda}) = h\mu/(\mu-\bar{\lambda})^2$, by (22), results in

$$d\bar{\phi}/d\mu = h\bar{\lambda}/(\mu-\bar{\lambda})^2 > 0 \qquad (\mu^*<\mu<\mu^{(1)}). \tag{25}$$

In a similar fashion we derive

$$d\bar{\phi}/dh = -\bar{\lambda}/(\mu-\bar{\lambda}) < 0 \qquad (h^{(1)}<h<h^*). \tag{26}$$

1.7. Example. $R(\lambda) = r\lambda$

Here

$$\phi(\lambda;\mu) = r\lambda - h\lambda/(\mu-\lambda) \qquad (\lambda<\mu). \tag{27}$$

$R''(\lambda) \equiv 0$, so Conditions 1 and 2 are met by all μ. By (20), (21), (22), $\mu^* = h/r$; $\bar{\lambda} = 0$ if $\mu \leq \mu^*$; $r = h\mu/(\mu-\bar{\lambda})^2$ if $\mu \geq \mu^*$. From the last equation an explicit solution can be obtained, namely

$$\bar{\lambda} = \mu - \sqrt{\mu h/r} \qquad (\mu \geq \mu^* = h/r). \tag{28}$$

This expression has also been obtained by Balachandran and Schaefer [to appear in IJPR]. Clearly, $\bar{\phi} = 0$ if $\mu \leq \mu^*$. Otherwise, $\bar{\phi}$ is calculated by substituting $\lambda = \bar{\lambda}$ from (28) into (27), the result being

$$\bar{\phi} = h + r\mu - 2\sqrt{hr\mu} \qquad (\mu \geq \mu^* = h/r), \tag{29}$$

showing, incidentally, that $\bar{\phi}$ depends on r and μ only through their product $r\mu$. Curiously, $\bar{\phi} = (\sqrt{r\mu} - \sqrt{h})^2$.

Our main interest here is how $\bar{\lambda}$ depends on the parameters μ, h, and r. First, by differentiation of (28) with respect to μ, we obtain

$$d\bar{\lambda}/d\mu = 1 - \frac{1}{2}\sqrt{h/(r\mu)} \qquad (\mu \geq \mu^* = h/r),$$

$$d^2\bar{\lambda}/d\mu^2 > 0 \qquad (\mu \geq \mu^* = h/r).$$

In the present case, then, $\bar{\lambda}$ is an increasing, strictly convex function on $[\mu^*,\infty)$. Interestingly, $d\bar{\lambda}/d\mu = 1/2$ at $\mu = \mu^*$, and $\lim_{\mu\to\infty} d\bar{\lambda}/d\mu = 1$.

Equation (28) can be rewritten

$$\frac{\bar{\lambda}}{\mu^*} = \frac{\mu}{\mu^*} - \sqrt{\frac{\mu}{\mu^*}} \qquad (\frac{\mu}{\mu^*} \geq 1).$$

Hence $\bar{\lambda}/\mu^*$ depends solely on the ratio μ/μ^*. This fact permits the construction of a normalized graph from which $\bar{\lambda}$ can be easily derived for any μ and $\mu^* = h/r$. $\bar{\lambda}/\mu^*$ is depicted as a function of μ/μ^* in Figure IV-1.

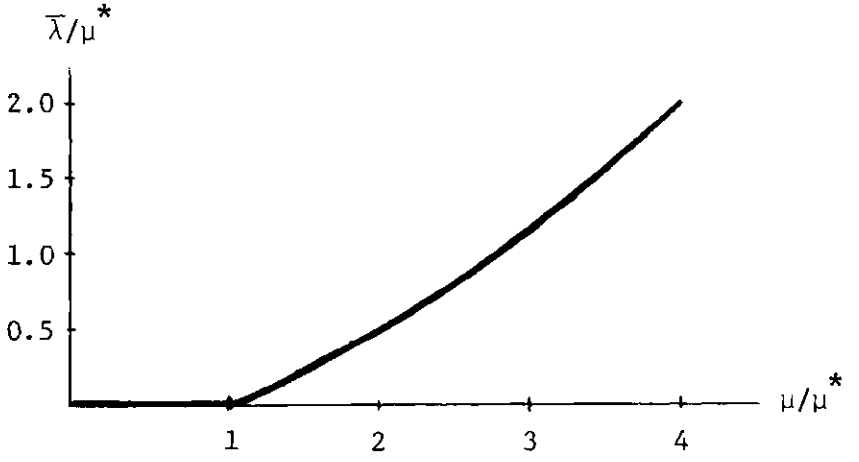

Figure IV-1. A Normalized Graph for $R(\lambda) = r\lambda$

By differentiation of (28) we derive easily

$$d\bar{\lambda}/dh = -\frac{1}{2}\sqrt{\mu/(rh)} < 0,$$

$$d\bar{\lambda}/dr = \frac{1}{2}\sqrt{\mu h/r^3} > 0.$$

We conclude this section with a simple numerical example we shall use throughout the chapter. Let $\mu = 1$, $R(\lambda) = r\lambda$, $r = 8$, $h = 2$. Note, $\mu^* = 0.25$. We find $\bar{\lambda} = 0.50$, $\bar{\phi} = 2.00$.

2. The Equilibrium Point Solution

Let $\bar{\lambda}_i(\mu)$ denote user i's optimal choice as a sole user, given that the service rate is μ. Let $\lambda_i[\lambda_j]$ be user i's optimal choice given the choice λ_j by user j (i=1,2), with the service rate μ being understood.

As explained, each two-user objective function is closely related to the corresponding one-user objective function, by (5). Hence, the functions $\lambda_i[\lambda_j]$ and $\bar{\lambda}_i(\mu)$ are closely related. Obviously, by (5), $\max_{\lambda_i} \pi_i(\lambda_i;\mu,\lambda_j) = \max_{\lambda_i} \phi_i(\lambda_i;\mu-\lambda_j)$, so

$$\lambda_i[\lambda_j] = \bar{\lambda}_i(\mu-\lambda_j) \qquad (i=1,2) \qquad (30)$$

with $\lambda_i[\lambda_j] = 0$ if $\mu-\lambda_j \leq 0$.

DEFINITION (Equilibrium Point). $(\hat{\lambda}_1,\hat{\lambda}_2)$ is said to be an equilibrium point if $\hat{\lambda}_1 = \lambda_1[\hat{\lambda}_2]$ and $\hat{\lambda}_2 = \lambda_2[\hat{\lambda}_1]$.

An equilibrium point is simply a pair of choices (λ_1,λ_2) such that none of the users can gain by making a different choice, provided the other user's choice remains the same. Clearly, behavioral assumptions made by users are a factor in determining the arrival rates ultimately selected by the two users of the service facility. We shall see that under some plausible behavioral assumptions, the users will settle for an equilibrium point solution if only the functions $\lambda_i[\lambda_j]$ (i=1,2) are sufficiently well behaved.

2.1. Dynamic Stability

Imagine that the decision-process evolves in time through a sequence of decisions on arrival rate made by the two users. Assume both arrival rates are known by both users. Suppose the users adjust their arrival

rates at certain intervals, with each user choosing his new arrival rate as some function of his and the other user's current arrival rates. The length of the time intervals between rate adjustments is immaterial to us since we are interested only in the final arrival rates.

A variety of adjustment processes are conceivable. Perhaps the simplest ones let each user make a complete adjustment to the current rate used by the other user, as if that rate will never be changed. In the real world it may seem unlikely that two users should adopt such a myopic strategy of decision-making. Yet in many markets a similar mode of decision-making has been frequently observed. Also, as we shall see, a reciprocal adjustment (or mutual accommodation) process as outlined may serve to effect a compromise between users with conflicting interests.

We shall consider two variants with complete adjustment, namely the alternate decision model, B_1, in which the users alternate adjusting their arrival rates, and the simultaneous decision model, B_2, in which the users adjust the arrival rates at the same time. Take an arbitrary initial arrival rate combination $(\lambda_1^0,\lambda_2^0)$. Let $(\lambda_1^n,\lambda_2^n)$, $n = 1,2,\ldots$, be the arrival rates in effect following the decision(s) at the n'th step. Under B_1, when user i makes the first move, we have: $(\lambda_i^1,\lambda_j^1) = (\lambda_i[\lambda_j^0],\lambda_j^0)$, $(\lambda_i^2,\lambda_j^2) = (\lambda_i^1,\lambda_j[\lambda_i^1])$, $(\lambda_i^3,\lambda_j^3) = (\lambda_i[\lambda_j^2],\lambda_j^2)$, etc. Under B_2: $(\lambda_i^n,\lambda_j^n) = (\lambda_i[\lambda_j^{n-1}], \lambda_j[\lambda_i^{n-1}])$, $n = 1,2,\ldots$.

DEFINITION (Dynamic Stability). An equilibrium point $(\hat{\lambda}_1,\hat{\lambda}_2)$ is said to be dynamically stable, if, for arbitrary $(\lambda_1^0,\lambda_2^0)$, $\lim_{n\to\infty} \lambda_i^n = \hat{\lambda}_i$, $i = 1,2$, under both B_1 and B_2.

LEMMA 7. Let $\lambda_1[\ ]$ and $\lambda_2[\ ]$ be continuous on $[0,\infty)$ such that

$-1 < d\lambda_i/d\lambda_j \leq 0$, $i = 1,2$. Then an equilibrium point $(\hat{\lambda}_1, \hat{\lambda}_2)$ exists, it is unique, and it is dynamically stable.

PROOF. In a Cartesian coordinate system, plot λ_1 and $\lambda_1[\lambda_2]$ along the x-axis, and plot λ_2 and $\lambda_2[\lambda_1]$ along the y-axis. Each point of intersection of the two curves representing $\lambda_1[\ ]$ and $\lambda_2[\ ]$ is an equilibrium point. Obviously, the curves intersect once and only once given the slope of the curves. A demonstration of dynamic stability as defined is also quite straightforward, but is rather tedious since various classes of starting points need to be considered under both B_1 and B_2. We omit the proof. □

Now return to our two-user model. Assume Condition 2 is met by both users. Then, by our results in Section 1, $\bar{\lambda}_i(\ )$ is continuous on $(0,\mu]$, and $0 \leq d\bar{\lambda}_i/dm < 1$ for each service rate $m \in (0,\mu]$, $i = 1,2$. By (30)

$$d\lambda_i/d\lambda_j = - d\bar{\lambda}_i/dm\big|_{m=\mu-\lambda_j} \qquad (0 \leq \lambda_j < \mu). \tag{31}$$

Consequently,

$$-1 < d\lambda_i/d\lambda_j \leq 0 \qquad (i=1,2). \tag{32}$$

Hence, if both users meet Condition 2, then the conditions of Lemma 7 are satisfied. Applying the lemma we obtain

THEOREM 2. Let both users satisfy Condition 2, that is, $R_i''(\lambda_i) < h_i/(\mu-\lambda_i)^2$ for all $\lambda_i < \mu$, $i = 1,2$. Then an equilibrium point $(\hat{\lambda}_1, \hat{\lambda}_2)$ exists, it is unique, and it is dynamically stable.

2.2. Equilibrium Point Solution and One-User Solutions

Suppose Condition 2 is met by both users. In general, the equilibrium point solution $(\hat{\lambda}_1,\hat{\lambda}_2)$ has to be calculated by numerical methods. However, we are interested not only in $(\hat{\lambda}_1,\hat{\lambda}_2)$ and the corresponding profits $(\hat{\pi}_1,\hat{\pi}_2)$, but also in how these quantities compare with $(\overline{\lambda}_1,\overline{\lambda}_2)$ and $(\overline{\phi}_1,\overline{\phi}_2)$, respectively. Some simple relationships exist.

There are exactly four different cases to consider. We label them (a), (b), (c), and (d). All four cases are illustrated in Figure IV-2. The equilibrium point solution $(\hat{\lambda}_1,\hat{\lambda}_2)$ is marked by a circle. Note, $\overline{\lambda}_i = \lambda_i[0]$ (i=1,2). The origin is (0,0).

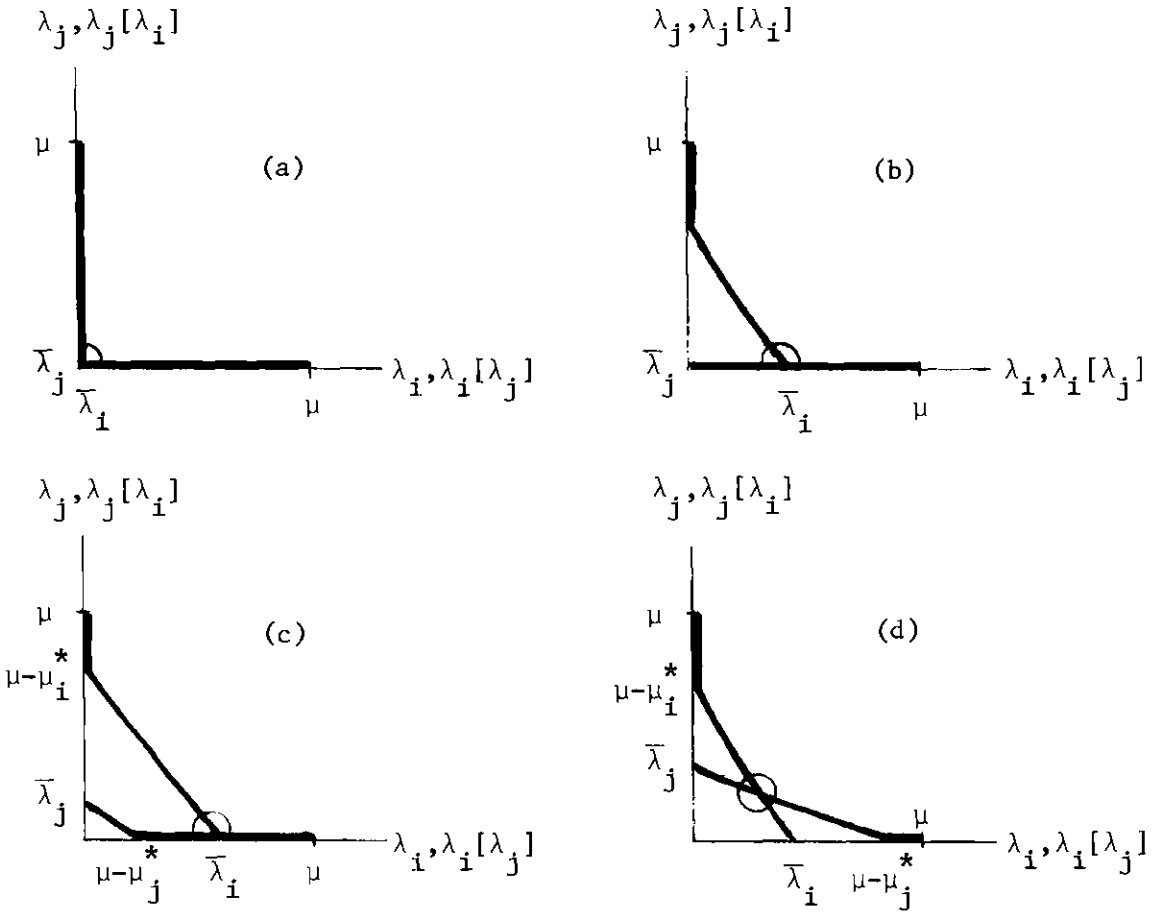

Figure IV-2. $(\hat{\lambda}_1,\hat{\lambda}_2)$ in the Four Possible Cases.

Using Figure IV-2 for reference one can convince himself of the truth of the equivalence relations (33)-(36) below. First of all, the four cases (a)-(d) as defined by the left-hand side of the equivalence relations are mutually exclusive and together they exhaust all possibilities. In order to see this, observe that when $\overline{\lambda}_i > 0$ and $\overline{\lambda}_j > 0$, then $\overline{\lambda}_i \geq \mu - \mu_j^*$ implies $\overline{\lambda}_j < \mu - \mu_i^*$, due to the slope of the curves.

Since Condition 2 is met, Equation (32) holds and, by Theorem 2, $(\hat{\lambda}_1, \hat{\lambda}_2)$ is the unique equilibrium point. Using these facts in conjunction with Equations (1) and (7) the right-hand sides of (33)-(36) are easily derived. Perhaps the only results that are not evident are $\hat{\pi}_j < \overline{\phi}_j$ in case (c), and $\hat{\pi}_i < \overline{\phi}_i$ $(i=1,2)$ in case (d). We have: (c) $\hat{\pi}_j \equiv \pi_j(0;\mu,\hat{\lambda}_i) = \phi_j(0;\mu) < \phi_j(\overline{\lambda}_j;\mu) \equiv \overline{\phi}_j$; (d) $\hat{\pi}_i \equiv \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j) < \pi_i(\hat{\lambda}_i;\mu,0) = \phi_i(\hat{\lambda}_i;\mu) < \phi_i(\overline{\lambda}_i;\mu) \equiv \overline{\phi}_i$.

$$\text{(a)} \qquad \left\langle \overline{\lambda}_i = 0, \overline{\lambda}_j = 0 \right\rangle \Leftrightarrow \left\langle (0 = \hat{\lambda}_i = \overline{\lambda}_i = \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_i = \overline{\phi}_i),\ i=1,2 \right\rangle . \tag{33}$$

$$\text{(b)} \qquad \left\langle \overline{\lambda}_i > 0, \overline{\lambda}_j = 0 \right\rangle \Leftrightarrow \left\langle \begin{matrix} 0 < \hat{\lambda}_i = \overline{\lambda}_i = \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_i = \overline{\phi}_i, \\ 0 = \hat{\lambda}_j = \overline{\lambda}_j < \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_j = \overline{\phi}_j \end{matrix} \right\rangle . \tag{34}$$

$$\text{(c)} \qquad \left\langle \overline{\lambda}_i > 0, \overline{\lambda}_j > 0, \overline{\lambda}_i \geq \mu - \mu_j^* \right\rangle \Leftrightarrow \left\langle \begin{matrix} 0 < \hat{\lambda}_i = \overline{\lambda}_i = \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_i = \overline{\phi}_i, \\ 0 = \hat{\lambda}_j < \overline{\lambda}_j < \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_j < \overline{\phi}_j \end{matrix} \right\rangle . \tag{35}$$

$$\text{(d)} \qquad \left\langle 0 < \overline{\lambda}_i < \mu - \mu_j^*,\ i=1,2 \right\rangle \Leftrightarrow \left\langle (0 < \hat{\lambda}_i < \overline{\lambda}_i < \hat{\lambda}_1 + \hat{\lambda}_2,\ \hat{\pi}_i < \overline{\phi}_i),\ i=1,2 \right\rangle . \tag{36}$$

From the right-hand sides of the equivalence relations (33)-(36) we obtain easily the conclusion of our next theorem.

THEOREM 3. Let both users satisfy Condition 2. Then $0 \leq \hat{\lambda}_i \leq \overline{\lambda}_i \leq \hat{\lambda}_1 + \hat{\lambda}_2$ and $\hat{\pi}_i \leq \overline{\phi}_i$ for $i = 1,2$.

In other words, if Condition 2 is met, and the users follow strategies that lead to the choices $(\hat{\lambda}_1,\hat{\lambda}_2)$, then each user's arrival rate will be less than or equal to the arrival rate he would choose if alone, but the aggregate arrival rate will be greater than or equal to both of the one-user rates. Also, obviously, the profit will decrease or at best stay the same in the event of the introduction of a second user.

2.3. Sensitivity Analysis. Preliminaries

The last question to be addressed is how $(\hat{\lambda}_1,\hat{\lambda}_2)$ and associated profits $(\hat{\pi}_1,\hat{\pi}_2)$ are affected by small changes in the parameters μ, h_1 and h_2. Again, only parameter sets which satisfy Condition 2 are considered.

In cases (a), (b), and (c), $\hat{\lambda}_i$ equals 0 or $\overline{\lambda}_i$ (i=1,2). Thus it is a simple matter to deduce the effect of a change in μ, h_1 or h_2. We just apply the results of the analysis of the one-user model, disregarding the inactive user.

Case (d) is the most interesting case. Here, for the given R_1, R_2, μ, h_1, h_2, we have $0 < \hat{\lambda}_1 < \mu$, $0 < \hat{\lambda}_2 < \mu$. Let α_1 and α_2 denote the slopes of the two curves at $\hat{\lambda}_2$ and $\hat{\lambda}_1$, respectively. That is,

$$\alpha_i = d\lambda_i[\lambda_j]/d\lambda_j\big|_{\lambda_j=\hat{\lambda}_j} \qquad (i=1,2). \tag{37}$$

By (31)

$$\alpha_i = -\, d\overline{\lambda}_i/dm\big|_{m=\mu-\hat{\lambda}_j} \qquad (i=1,2). \tag{38}$$

α_1 and α_2 may be evaluated by use of (23), provided $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are already known. Observe, $-1 < \alpha_i < 0$, $i = 1,2$.

The two intersecting curves represent functions that will be

considered dependent on a parameter p. A small change in p will result in a small change in one or both functions. To us, what matters is the rate of change in the two functions, at $\hat{\lambda}_2$ and $\hat{\lambda}_1$, respectively. Let

$$\beta_i = d\lambda_i[\lambda_j]/dp\big|_{\lambda_j=\hat{\lambda}_j} \qquad (i=1,2). \tag{39}$$

By standard methods of calculus we derive

$$d\hat{\lambda}_i/dp = (\beta_i+\beta_j\alpha_i)/(1-\alpha_1\alpha_2) \quad (i=1,2). \tag{40}$$

2.4. Effects of a Change in μ

First we discuss the case p = μ. An increase in μ by $\Delta\mu > 0$ will simply result in a translation of the two curves away from the origin. Referring to Figure IV-2(d), $\lambda_i[\ ]$ will slide along the y-axis, $\lambda_j[\ ]$ along the x-axis, both curves being displaced by Δμ units. It is seen that here $\beta_1 = -\alpha_1$ and $\beta_2 = -\alpha_2$. Inserting into (40) we obtain

$$d\hat{\lambda}_i/d\mu = (-\alpha_i)(1+\alpha_j)/(1-\alpha_1\alpha_2) \quad (i=1,2). \tag{41}$$

Now, as $-1 < \alpha_i < 0$ (i=1,2), we have $0 < -\alpha_i < 1$, $0 < 1+\alpha_i < 1$, $0 < 1-\alpha_1\alpha_2 < 1$ (i=1,2). We conclude

$$0 < d\hat{\lambda}_i/d\mu < 1 \qquad (i=1,2).$$

Also,

$$0 < d(\hat{\lambda}_1+\hat{\lambda}_2)/d\mu < 1.$$

One might expect both users to increase the profit when, in case (d), μ is increased. This we shall prove. Let $\Delta\hat{\pi}_i$ denote the change in profit for user i caused by an increase Δμ in the service rate. We must show $\Delta\hat{\pi}_i > 0$ if $\Delta\mu > 0$.

By definition,

$$\Delta\hat{\pi}_i = \pi_i(\hat{\lambda}_i+\Delta\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j) - \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j).$$

This may be rewritten as

$$\begin{aligned}\Delta\hat{\pi}_i = {} & [\pi_i(\hat{\lambda}_i+\Delta\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j) - \pi_i(\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j)] \\ & + [\pi_i(\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j) - \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j)].\end{aligned} \tag{42}$$

By definition of an equilibrium point and the fact $\Delta\hat{\lambda}_i > 0$, clearly

$$\pi_i(\hat{\lambda}_i+\Delta\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j) - \pi_i(\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j) > 0. \tag{43}$$

Additionally,

$$\begin{aligned}\pi_i(\hat{\lambda}_i;\mu+\Delta\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j)-\pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j) &= \phi_i(\hat{\lambda}_i;(\mu-\hat{\lambda}_j)+(\Delta\mu-\Delta\hat{\lambda}_j))-\phi_i(\hat{\lambda}_i;\mu-\hat{\lambda}_j) \\ &> 0 \quad [\hat{\lambda}_i>0;\text{Eq.}(6);d\hat{\lambda}_j/d\mu<1 \Rightarrow \Delta\hat{\lambda}_j<\Delta\mu].\end{aligned} \tag{44}$$

By (42), (43), and (44), $\Delta\hat{\pi}_i > 0$. By symmetry, $\Delta\hat{\pi}_j > 0$.

2.5. Effects of a Change in h_1 or h_2

Next we discuss the case $p = h_i$ $(i=1,2)$. Take $p = h_i$. Here $\beta_i = d\bar{\lambda}_i/dh_i$, evaluated by (24) with $\mu-\hat{\lambda}_j$ replacing μ, and $\hat{\lambda}_i$ replacing $\bar{\lambda}_i$ (subscript i added in (24)). By (24), $\beta_i < 0$. Also, $\beta_j = d\bar{\lambda}_j/dh_i = 0$. Thus, by (40),

$$d\hat{\lambda}_i/dh_i = \beta_i/(1-\alpha_1\alpha_2) < 0, \tag{45}$$

$$d\hat{\lambda}_j/dh_i = \alpha_j\beta_i/(1-\alpha_1\alpha_2) > 0. \tag{46}$$

Also worth noting,

$$d(\hat{\lambda}_1+\hat{\lambda}_2)/dh_i = \beta_i(1+\alpha_j)/(1-\alpha_1\alpha_2) < 0. \tag{47}$$

The direction of change in profits is found by a series of comparisons. Since h_i has not been included as an argument of π_i we shall employ the usual "conditional upon" or "given" notation. Let $\Delta h_i > 0$. Recall that in case (d) under discussion, $\hat{\lambda}_1 > 0$, $\hat{\lambda}_2 > 0$. By (45), (46), then we have $\Delta\hat{\lambda}_i < 0$, $\Delta\hat{\lambda}_j > 0$. We may assume Δh_i is small enough so that $\hat{\lambda}_i + \Delta\hat{\lambda}_i > 0$.

First we examine the effect on user i's profit. We have

$$\begin{aligned}
\hat{\pi}_i|_{h_i} &\equiv \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j)|_{h_i} > \pi_i(\hat{\lambda}_i+\Delta\hat{\lambda}_i;\mu,\hat{\lambda}_j)|_{h_i} && \text{[by definition of } (\hat{\lambda}_1,\hat{\lambda}_2)] \\
&> \pi_i(\hat{\lambda}_i+\Delta\hat{\lambda}_i;\mu,\hat{\lambda}_j+\Delta\hat{\lambda}_j)|_{h_i+\Delta h_i} && \text{[by (1)]} \\
&\equiv \hat{\pi}_i|_{h_i+\Delta h_i}.
\end{aligned}$$

Thus, $\Delta\hat{\pi}_i = \hat{\pi}_i|_{h_i+\Delta h_i} - \hat{\pi}_i|_{h_i} < 0$ if $\Delta h_i > 0$.

Next we examine the effect on user j's profit. We have

$$\begin{aligned}
\hat{\pi}_j|_{h_i} &\equiv \pi_j(\hat{\lambda}_j;\mu,\hat{\lambda}_i)|_{h_i} < \pi_j(\hat{\lambda}_j;\mu,\hat{\lambda}_i+\Delta\hat{\lambda}_i)|_{h_i} && \text{[by (1)]} \\
&= \pi_j(\hat{\lambda}_j;\mu,\hat{\lambda}_i+\Delta\hat{\lambda}_i)|_{h_i+\Delta h_i} \\
&\leq \max_{\lambda_j} \pi_j(\lambda_j;\mu,\hat{\lambda}_i+\Delta\hat{\lambda}_i)|_{h_i+\Delta h_i} \\
&= \pi_j(\hat{\lambda}_j+\Delta\hat{\lambda}_j;\mu,\hat{\lambda}_i+\Delta\hat{\lambda}_i)|_{h_i+\Delta h_i} \\
&\equiv \hat{\pi}_j|_{h_i+\Delta h_i}.
\end{aligned}$$

Thus, $\Delta\hat{\pi}_j = \hat{\pi}_j|_{h_i+\Delta h_i} - \hat{\pi}_j|_{h_i} > 0$ if $\Delta h_i > 0$. Because of the symmetry in case (d), the results apply for i = 1,2.

Our results are summarized in Theorem 4 that covers all cases, not just case (d). For the other cases the results follow easily from previous

remarks and the findings in Section 1.

THEOREM 4. Let both users satisfy Condition 2 for given R_1, R_2, μ, h_1, h_2. Then the equilibrium point solution $(\hat{\lambda}_1,\hat{\lambda}_2)$ has the properties

$$\text{(i)} \qquad 0 \leq d\hat{\lambda}_i/d\mu < 1 \qquad (i=1,2),$$

$$\text{(ii)} \qquad 0 \leq d(\hat{\lambda}_1+\hat{\lambda}_2)/d\mu < 1,$$

$$\text{(iii)} \qquad d\hat{\pi}_i/d\mu \geq 0 \qquad (i=1,2).$$

Equality holds in (i) and (iii) if and only if $\hat{\lambda}_i = 0$; and in (ii) if and only if $\hat{\lambda}_1 = \hat{\lambda}_2 = 0$. Also,

$$\text{(iv)} \qquad d\hat{\lambda}_i/dh_i \leq 0,\ d\hat{\lambda}_j/dh_i \geq 0 \qquad (i=1,2),$$

$$\text{(v)} \qquad d(\hat{\lambda}_1+\hat{\lambda}_2)/dh_i \leq 0 \qquad (i=1,2),$$

$$\text{(vi)} \qquad d\hat{\pi}_i/dh_i \leq 0,\ d\hat{\pi}_j/dh_i \geq 0 \qquad (i=1,2).$$

Equality holds in first statement of (iv) and (vi) if and only if $\hat{\lambda}_i = 0$; in second statement of (iv) and (vi) if and only if $\hat{\lambda}_i = 0$ or $\hat{\lambda}_j = 0$; and in (v) if and only if $\hat{\lambda}_i = 0$.

(i) and (ii) become evident when one realizes that a translation $\Delta\mu$ of curves with slopes greater than -1 but nonpositive (by Condition 2) will cause $(\hat{\lambda}_1+\Delta\hat{\lambda}_1, \lambda_2+\Delta\hat{\lambda}_2)$ to be located in the triangle defined by $(\hat{\lambda}_1,\hat{\lambda}_2)$, $(\hat{\lambda}_1+\Delta\mu,\hat{\lambda}_2)$, $(\hat{\lambda}_1,\hat{\lambda}_2+\Delta\mu)$, excepting the line connecting the last two points. (iv) is obvious given the slopes of the curves and the downward shift in $\lambda_i[\ ]$ for a $\Delta h_i > 0$. Now, for user j, a reduction $\Delta\hat{\lambda}_i < 0$ has the effect of an increase $-\Delta\hat{\lambda}_i$ in μ. Thus, under Condition 2, $0 \leq \Delta\hat{\lambda}_j < -\Delta\hat{\lambda}_i$. Hence, $\Delta\hat{\lambda}_i < 0$ implies that $\Delta\hat{\lambda}_i+\Delta\hat{\lambda}_j < 0$, explaining (v).

2.6. Example. $R_1(\lambda_1) = r_1\lambda_1$, $R_2(\lambda_2) = r_2\lambda_2$

We begin with the observation that Condition 2 is met for all μ for both users. As explained, $(\hat{\lambda}_1,\hat{\lambda}_2)$ and $(\hat{\mu}_1,\hat{\mu}_2)$ are functions of

r_1, r_2, μ, h_1, h_2. For parameters such that case (a), (b), or (c), applies it is a simple matter to derive $(\hat{\lambda}_1,\hat{\lambda}_2)$ and $(\hat{\pi}_1,\hat{\pi}_2)$ by resort to the results of our one-user model analysis. Also, these quantities' dependence on μ, h_1, and h_2, can be determined in this way. Therefore, we will discuss only case (d), where the parameter values result in $\hat{\lambda}_1 > 0$, $\hat{\lambda}_2 > 0$. The discussion will be limited to the determination of $\hat{\lambda}_1$ and $\hat{\lambda}_2$.

First, we derive a useful necessary and sufficient condition for case (d), in our linear model. The basic observation is

$$\langle 0 < \overline{\lambda}_i < \mu-\mu_j^* \rangle \Leftrightarrow \langle \mu > \mu_i^*,\ \mu > (\mu_j^*)^2/\mu_i^* \rangle \qquad (i=1,2). \tag{48}$$

The variables are those defined for the one-user model. Recall that $\mu_i^* = h_i/r_i$. Assume $0 < \overline{\lambda}_i < \mu-\mu_j^*$. According to Section 1.7, $\overline{\lambda}_i > 0 \Rightarrow \mu > \mu_i^*$, and $\overline{\lambda}_i = \mu - \sqrt{\mu\mu_i^*}$. Substituting the expression for $\overline{\lambda}_i$ into $\overline{\lambda}_i < \mu-\mu_j^*$, and reducing, we derive $\mu > (\mu_j^*)^2/\mu_i^*$. Conversely, assume $\mu > \mu_i^*$ and $\mu > (\mu_j^*)^2/\mu_i^*$. By Section 1.7, $\mu > \mu_i^* \Rightarrow \overline{\lambda}_i > 0$, and, by (28), $\overline{\lambda}_i = \mu - \sqrt{\mu\mu_i^*}$. Now, $\mu > (\mu_j^*)^2/\mu_i^*$ is the same as $\sqrt{\mu\mu_i^*} > \mu_j^*$. Hence $\overline{\lambda}_i = \mu - \sqrt{\mu\mu_i^*} < \mu-\mu_j^*$. This proves (48).

Suppose $\mu_j^* \geq \mu_i^*$. Then $(\mu_j^*)^2/\mu_i^* \geq \mu_j^* \geq \mu_i^* \geq (\mu_i^*)^2/\mu_j^*$. Consequently,

$$\langle (\mu > \mu_i^*,\ \mu > (\mu_j^*)^2/\mu_i^*),\ i=1,2 \rangle \Leftrightarrow \mu > \frac{(\max(\mu_1^*,\mu_2^*))^2}{\min(\mu_1^*,\mu_2^*)}. \tag{49}$$

From (48) and (49) we conclude:

$$\text{case (d)} \equiv \langle (0 < \overline{\lambda}_i < \mu-\mu_j^*),\ i=1,2 \rangle \Leftrightarrow \mu > \frac{(\max(\mu_1^*,\mu_2^*))^2}{\min(\mu_1^*,\mu_2^*)}.$$

If $\mu_1^* = \mu_2^* = \mu^*$, then case (d) $\Leftrightarrow \mu > \mu^*$.

Assume the parameters meet the conditions for case (d). As always, $\hat{\lambda}_1 < \mu$, $\hat{\lambda}_2 < \mu$, $\hat{\lambda}_1 + \hat{\lambda}_2 < \mu$. In Section 1 we derived $\overline{\lambda} = \mu - \sqrt{\mu\mu^*}$ if $\mu > \mu^*$. Similarly, in the two-user model, $\lambda_i[\lambda_j] = (\mu - \lambda_j) - \sqrt{(\mu-\lambda_j)\mu_i^*}$ if $\mu - \lambda_j > \mu_i^*$, $i = 1,2$. In case (d) clearly $\mu - \hat{\lambda}_j > \mu_i^*$, $i = 1,2$, as seen from Figure IV-2(d). Hence,

$$\hat{\lambda}_i = \mu - \hat{\lambda}_j - \sqrt{(\mu-\hat{\lambda}_j)\mu_i^*} \qquad (i=1,2). \tag{50}$$

We see that $\sqrt{(\mu-\hat{\lambda}_2)\mu_1^*} = \sqrt{(\mu-\hat{\lambda}_1)\mu_2^*}$ by which

$$\frac{\mu-\hat{\lambda}_2}{\mu-\hat{\lambda}_1} = \frac{\mu_2^*}{\mu_1^*}.$$

Use of this relation in (50) results in

$$\hat{\lambda}_i = \frac{\mu_j^*}{\mu_i^*}(\mu-\hat{\lambda}_i) - \sqrt{(\mu-\hat{\lambda}_i)\mu_j^*} \qquad (i=1,2). \tag{51}$$

(51) is a quadratic equation in $\hat{\lambda}_i$. We shall not give the solution since the expression is rather complicated. In the special case in which the users are identical, $r_1 = r_2$ and $h_1 = h_2$, so that $\mu_1^* = \mu_2^* = \mu^*$, (51) reduces to

$$\hat{\lambda}_i = \mu - \hat{\lambda}_i - \sqrt{(\mu-\hat{\lambda}_i)\mu^*} \qquad (i=1,2).$$

Solving, we obtain

$$\hat{\lambda}_1 = \hat{\lambda}_2 = [4\mu - \mu^* - \sqrt{8\mu\mu^* + (\mu^*)^2}]/8.$$

In our numerical example with $\mu = 1$, $r_1 = r_2 = 8$, $h_1 = h_2 = 2$, we find $\hat{\lambda}_1 = \hat{\lambda}_2 = 0.29$, $\hat{\pi}_1 = \hat{\pi}_2 = 0.94$.

3. The Leader-Follower Solution

The model to be explored in this section is called the leader-follower model. In contrast with the previously analyzed two-user model in which both users are followers, now one user plays the role of a leader while the other is a follower. That is, the follower, say user j, automatically chooses the conditionally optimal arrival rate $\lambda_j[\lambda_i]$ for any choice λ_i by the leader, who will choose λ_i so as to maximize his profit, taking into account the follower's response $\lambda_j[\lambda_i]$.

New notation is needed to distinguish leader and follower. As usual, subscript i or j, shall denote the user, but we may add the number 1 to designate the leader, or 2 to designate the follower. Thus π_{i1} shall denote user i's objective function when he is the leader; λ^*_{i1} will be user i's optimal choice as a leader; $\lambda^*_{j2} \equiv \lambda_j[\lambda^*_{i1}]$ is the corresponding choice by user j as the follower; $\pi^*_{i1} \equiv \pi_i(\lambda^*_{i1};\mu,\lambda^*_{j2})$ and $\pi^*_{j2} \equiv \pi_j(\lambda^*_{j2};\mu,\lambda^*_{i1})$ denote the resulting profits.

If $\lambda_i < \mu$, then always $\lambda_j[\lambda_i] < \mu-\lambda_i$, or $\lambda_i + \lambda_j[\lambda_i] < \mu$. Hence, by (1), the leader's objective function is

$$\pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i]) = \begin{cases} R_i(\lambda_i)-h_i\lambda_i/(\mu-\lambda_i-\lambda_j[\lambda_i]) & (\lambda_i<\mu), \\ -\infty & (\lambda_{i=}\geq\mu), \end{cases} \quad (i=1,2). \quad (52)$$

Obviously, $\lambda^*_{i1} < \mu$, $i = 1,2$. If user i is the leader, then, of course, he may select $\lambda_i = \hat{\lambda}_i$ in which case the follower response is $\lambda_j = \lambda_j[\hat{\lambda}_i] = \hat{\lambda}_j$. Thus, a leader can assure himself of the equilibrium point profit, here

$\hat{\pi}_i = \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j)$. The question is whether he can do better. As we shall see, the answer is that frequently he can.

Throughout the analysis of this model R_1, R_2, μ, h_1, and h_2, are fixed, and we assume that Condition 2 is satisfied by both users.

3.1. $\hat{\lambda}_i \leq \lambda^*_{i1}$ and Related Results

We will show that the leader's optimal arrival rate is at least as great as his equilibrium arrival rate. Suppose user i is the leader. If $\hat{\lambda}_i = 0$, then, trivially, $\lambda^*_{i1} \geq \hat{\lambda}_i$, so assume $\hat{\lambda}_i > 0$. Then

$$\begin{aligned}
\hat{\pi}_i &\equiv \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j) \\
&> \pi_i(\lambda_i;\mu,\hat{\lambda}_j) \qquad (\lambda_i<\hat{\lambda}_i) \qquad [\text{by definition of } (\hat{\lambda}_1,\hat{\lambda}_2)] \\
&\geq \pi_i(\lambda_i;\mu,\lambda_j[\lambda_i]) \qquad (\lambda_i<\hat{\lambda}_i) \qquad [\lambda_j[\lambda_i] \geq \hat{\lambda}_j;\ \text{Eq.(1)}] \\
&\equiv \pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i]) .
\end{aligned}$$

Now, $\hat{\pi}_i > \pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i])$ for all $\lambda_i < \hat{\lambda}_i$ implies $\lambda^*_{i1} \geq \hat{\lambda}_i$. We conclude that, in any case,

$$\hat{\lambda}_i \leq \lambda^*_{i1} \qquad (i=1,2) .$$

Correspondingly, $\lambda^*_{j2} \leq \hat{\lambda}_j$.

Evidently, $\pi^*_{i1} \geq \hat{\pi}_i$. On the other hand, $\pi^*_{j2} \leq \hat{\pi}_j$. To see this, note that

$$\pi_j(\lambda_j;\mu,\lambda^*_{i1}) \leq \pi_j(\lambda_j;\mu,\hat{\lambda}_i) \qquad (\text{all } \lambda_j) \qquad [\lambda^*_{i1} \geq \hat{\lambda}_i;\ \text{Eq.(1)}],$$

which implies

$$\pi^*_{j2} = \max_{\lambda_j} \pi_j(\lambda_j;\mu,\lambda^*_{i1}) \leq \max_{\lambda_j} \pi_j(\lambda_j;\mu,\hat{\lambda}_i) = \hat{\pi}_j.$$

The above inequalities can be strengthened somewhat in special cases. It is easy to see that in cases (a) and (b), and also in case (c), provided the leader is user i, where $\overline{\lambda}_i \geq \mu-\mu^*_j$, we have $(\lambda^*_{i1},\lambda^*_{j2}) = (\hat{\lambda}_i,\hat{\lambda}_j)$. Consequently, equality holds in these cases, and the leader-follower solution is identical to the equilibrium point solution. Also in case (d), to be discussed next, can the inequalities be strengthened.

3.2. Case (d): $\hat{\lambda}_i < \lambda^*_{i1} \leq \mu-\mu^*_j$

It has been shown that, in general, $\hat{\lambda}_i \leq \lambda^*_{i1}$. We will show that in case (d) $\hat{\lambda}_i < \lambda^*_{i1} \leq \mu-\mu^*_j$. First we prove $\hat{\lambda}_i < \lambda^*_{i1}$.

Note, π_{i1} as given by (52) is continuous and right differentiable with respect to λ_i on $[0,\mu)$. (Left and right derivatives are different at $\lambda_i=\mu-\mu^*_j$, so we focus on the right derivative.) Differentiation of (52) results in the following expression for the right derivative:

$$\frac{d\pi_{i1}}{d\lambda_i} = R_i'(\lambda_i) - h_i \frac{\mu-(-\lambda_j'[\lambda_i])\lambda_i-\lambda_j[\lambda_i]}{(\mu-\lambda_i-\lambda_j[\lambda_i])^2} \qquad (\lambda_i<\mu). \tag{53}$$

In (53), $\lambda_j'[\lambda_i] = d\lambda_j[\lambda_i]/d\lambda_i$. For $\lambda_i < \mu-\mu^*_j$ we have $\lambda_j[\lambda_i] > 0$ and $-\lambda_j'[\lambda_i] > 0$, but for $\lambda_i \geq \mu-\mu^*_j$ we have $\lambda_j[\lambda_i] = 0$ and $-\lambda_j'[\lambda_i] = 0$.

Recall that in case (d) $\hat{\lambda}_1 > 0$, $\hat{\lambda}_2 > 0$. Hence, by definition of $(\hat{\lambda}_1,\hat{\lambda}_2)$, $d\pi_i/d\lambda_i\big|_{\lambda_i=\hat{\lambda}_i,\lambda_j=\hat{\lambda}_j} = 0$. By (1) then

$$R_i'(\hat{\lambda}_i) - h_i \frac{\mu-\hat{\lambda}_j}{(\mu-\hat{\lambda}_1-\hat{\lambda}_2)^2} = 0. \tag{54}$$

From (53) and (54)

$$\frac{d\pi_{il}}{d\lambda_i}\Big|_{\lambda_i=\hat{\lambda}_i} = R_i'(\hat{\lambda}_i) - h_i \frac{\mu-(-\lambda_j'[\hat{\lambda}_i])\hat{\lambda}_i-\hat{\lambda}_j}{(\mu-\hat{\lambda}_1-\hat{\lambda}_2)^2} \qquad \text{[by (53)]}$$

$$> R_i'(\hat{\lambda}_i) - h_i \frac{\mu-\hat{\lambda}_j}{(\mu-\hat{\lambda}_1-\hat{\lambda}_2)^2} \qquad [-\lambda_j'[\hat{\lambda}_i] > 0,\ \hat{\lambda}_i > 0]$$

$$= 0 . \qquad \text{[by (54)]}$$

Clearly, $d\pi_{il}/d\lambda_i|_{\lambda_i=\hat{\lambda}_i} > 0$ implies $\hat{\lambda}_i \neq \hat{\lambda}_{il}^*$. It is known already that $\hat{\lambda}_i \leq \lambda_{il}^*$. We conclude $\hat{\lambda}_i < \lambda_{il}^*$.

Next we prove $\lambda_{il}^* \leq \mu-\mu_j^*$ by showing that π_{il} is strictly decreasing on $[\mu-\mu_j^*,\mu)$. By remarks above, (53) specializes to

$$d\pi_{il}/d\lambda_i = R_i'(\lambda_i)-h_i\mu/(\mu-\lambda_i)^2 \qquad (\mu-\mu_j^* \leq \lambda_i < \mu). \qquad (55)$$

In case (d), $\bar{\lambda}_i > 0$ and satisfies $\phi_i'(\bar{\lambda}_i;\mu)=0$. Note, Condition 1, which is implied by our Condition 2, imposes strict concavity on ϕ_i. Hence $\phi_i'(\lambda_i;\mu) < \phi_i'(\bar{\lambda}_i;\mu)$ for all $\lambda_i > \bar{\lambda}_i$. It follows that $\phi_i'(\lambda_i;\mu) < 0$ for all $\lambda_i > \bar{\lambda}_i$. That is, by (8),

$$d\phi_i(\lambda_i;\mu)/d\lambda_i = R_i'(\lambda_i)-h_i\mu/(\mu-\lambda_i)^2 < 0 \qquad (\lambda_i>\bar{\lambda}_i). \qquad (56)$$

Since in case (d) $\bar{\lambda}_i < \mu-\mu_j^*$, clearly $\lambda_i \geq \mu-\mu_j^*$ implies $\lambda_i > \bar{\lambda}_i$. By (55) and (56), then,

$$d\pi_{il}/d\lambda_i = d\phi_i(\lambda_i;\mu)/d\lambda_i < 0 \qquad (\mu-\mu_j^* \leq \lambda_i < \mu).$$

This proves that $\lambda_{il}^* \leq \mu-\mu_j^*$. In summary,

$$\hat{\lambda}_i < \lambda_{il}^* \leq \mu-\mu_j^* \quad \text{(case (d))}. \qquad (57)$$

From (57) we deduce easily, see Figure IV-2(d), that $0 \leq \lambda^*_{j2} < \hat{\lambda}_j$, with $\lambda^*_{j2} = 0$ if and only if $\lambda^*_{i1} = \mu - \mu^*_j$. If, instead, user i had been the follower and user j the leader, everything said holds with subscripts i and j interchanged. Thus we arrive at the statement

$$0 \leq \lambda^*_{i2} < \hat{\lambda}_i < \lambda^*_{i1} \leq \mu - \mu^*_j \quad (i=1,2) \quad \text{(case (d))}. \tag{58}$$

As for profits, obviously $\hat{\pi}_i < \pi^*_{i1}$. On the other hand, $\hat{\pi}_i > \pi^*_{i2}$ as

$$\begin{aligned}
\hat{\pi}_i &\equiv \pi_i(\hat{\lambda}_i;\mu,\hat{\lambda}_j) \\
&> \pi_i(\lambda^*_{i2};\mu,\hat{\lambda}_j) \qquad [\lambda^*_{i2} < \hat{\lambda}_i \text{ and definition of } (\hat{\lambda}_1,\hat{\lambda}_2)] \\
&\geq \pi_i(\lambda^*_{i2};\mu,\lambda^*_{j1}) \qquad [\lambda^*_{j1} > \hat{\lambda}_j, \text{ Eq. (1)}] \\
&\equiv \pi^*_{i2}.
\end{aligned}$$

Combining the last two inequalities we obtain

$$\pi^*_{i2} < \hat{\pi}_i < \pi^*_{i1} \quad (i=1,2) \quad \text{(case (d))}. \tag{59}$$

Finally, we note that the optimal aggregate arrival rate in the leader-follower model exceeds $\hat{\lambda}_1 + \hat{\lambda}_2$ in case (d). As $-1 < d\lambda_j[\lambda_i]/d\lambda_i < 0$ everywhere on $(\hat{\lambda}_i, \mu - \mu^*_j)$, it must be true that $\lambda^*_{i1} - \hat{\lambda}_i > \hat{\lambda}_j - \lambda^*_{j2}$. Hence,

$$\hat{\lambda}_1 + \hat{\lambda}_2 < \lambda^*_{i1} + \lambda^*_{j2} \quad (i=1,2) \quad \text{(case (d))}. \tag{60}$$

3.3. Statement of Results

We are now ready to state the results of the analysis. Our theorem covers all cases, (a)-(d). In cases (a) and (b) proofs are trivial and are omitted. In case (c) the proof has been left out in order not to

repeat arguments used in case (d).

THEOREM 5. Let both users satisfy Condition 2. Then $\lambda^*_{i2} \leq \hat{\lambda}_i \leq \lambda^*_{i1} \leq \max(\overline{\lambda}_i, \mu-\mu^*_j)$, $\hat{\lambda}_1 + \hat{\lambda}_2 \leq \lambda^*_{i1} + \lambda^*_{j2}$, $\pi^*_{i2} \leq \hat{\pi}_i \leq \pi^*_{i1}$, $i = 1,2$. Equality holds everywhere in cases (a) and (b). In case (d) strict inequality holds in all relations between λ's or π's.

We see that a user may gain and will never lose by assuming leadership rather than following the other user. In cases where it is to both users' advantage to lead rather than follow, as in case (d), the stage is set for conflict. In the real world a fight for supremacy may occur, or one user may resist another user's attempt to set himself up as a leader. The two users may prefer to settle for a compromise solution. One may view mutual accommodation as in the follower-follower models as a compromise. Another possible resolution of the conflict is cooperation which we shall study in the next section.

3.4. Example. $R_1(\lambda_1) = r_1\lambda_1$, $R_2(\lambda_2) = r_2\lambda_2$

Note, Condition 2 is met by both users. Suppose user i, i = 1 or 2, is the leader, and user j is the follower. If $\overline{\lambda}_i = 0$ or $\overline{\lambda}_i \geq \mu-\mu^*_j$, then clearly $\lambda^*_{i1} = \overline{\lambda}_i$. Assume therefore r_1, r_2, μ, h_1, h_2 are such that $0 < \overline{\lambda}_i < \mu-\mu^*_j$. Our sole objective is the determination of λ^*_{i1}. By (52)

$$\pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i]) = r_i\lambda_i - h_i\lambda_i/(\mu-\lambda_i-\lambda_j[\lambda_i]) \qquad (\lambda_i<\mu). \tag{61}$$

Consider the function π_{i1} on $[0,\mu-\mu^*_j]$. By remarks in Section 2, if $\lambda_i \leq \mu-\mu^*_j$, then $\lambda_j[\lambda_i] = \mu - \lambda_i - \sqrt{(\mu-\lambda_i)\mu^*_j}$. Insertion of this expression into (61) yields

$$\pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i]) = r_i\lambda_i - h_i\lambda_i/\sqrt{(\mu-\lambda_i)\mu_j^*} \qquad (\lambda_i \leq \mu-\mu_j^*). \tag{62}$$

Differentiation leads to

$$d\pi_{i1}/d\lambda_i = r_i - h_i\mu_j^*(\mu-\lambda_i/2)[(\mu-\lambda_i)\mu_j^*]^{-3/2}, \tag{63}$$

$$d^2\pi_{i1}/d\lambda_i^2 = -h_i(\mu_j^*)^2(\mu-\lambda_i/4)[(\mu-\lambda_i)\mu_j^*]^{-5/2}. \tag{64}$$

By (64), π_{i1} is strictly concave on $[0,\mu-\mu_j^*]$.

For $\lambda_i \in [\mu-\mu_j^*,\mu)$ we have $\pi_{i1}(\lambda_i;\mu,\lambda_j[\lambda_i]) = \phi_i(\lambda_i;\mu)$. By Condition 2 ϕ_i is strictly concave on $[0,\mu)$. Hence, π_{i1} is strictly concave also on $[\mu-\mu_j^*,\mu)$.

At $\lambda_i = \mu-\mu_j^*$ the left and right derivatives differ. By (63) and (8) we have

$$d\pi_{i1}/d\lambda_i\big|_{(\mu-\mu_j^*)-} = r_i - h_i\tfrac{1}{2}(\mu+\mu_j^*)/(\mu_j^*)^2, \tag{65}$$

$$d\pi_{i1}/d\lambda_i\big|_{(\mu-\mu_j^*)+} = r_i - h_i\mu/(\mu_j^*)^2. \tag{66}$$

It is seen that the left derivative is greater than the right derivative at $\lambda_i = \mu-\mu_j^*$. We conclude that π_{i1} is strictly concave on $[0,\mu)$ and therefore has a unique maximum.

By (48), $0 < \bar{\lambda}_i < \mu-\mu_j^*$ implies $\mu > (\mu_j^*)^2/\mu_i^*$, which in turn implies $d\pi_{i1}/d\lambda_i\big|_{(\mu-\mu_j^*)+} < 0$. Hence, $\lambda_{i1}^* \in [0,\mu-\mu_j^*]$. Perhaps λ_{i1}^* is most easily calculated directly from (62) by exploitation of the concavity property.

By (63), $d\pi_{i1}/d\lambda_i\big|_0 = r_i - h_i/\sqrt{\mu\mu_j^*}$. If $d\pi_{i1}/d\lambda_i\big|_0 \leq 0$, or, equivalently, $\mu\mu_j^*/(\mu_i^*)^2 \leq 1$, then $\lambda_{i1}^* = 0$. Similarly, if $d\pi_{i1}/d\lambda_i\big|_{(\mu-\mu_j^*)-} \geq 0$, or, equivalently, by (65), $\frac{1}{2}(\mu+\mu_j^*)\mu_i^*/(\mu_j^*)^2 \leq 1$, then $\lambda_{i1}^* = \mu-\mu_j^*$. Otherwise, $0 < \lambda_{i1}^* < \mu-\mu_j^*$. Observe that if $\mu_1^* = \mu_2^*$, then, always, $0 < \lambda_{i1}^* < \mu-\mu_j^*$.

In our numerical example with $\mu = 1$, $r_1 = r_2 = 8$, $h_1 = h_2 = 2$, we have $\bar{\lambda}_i = 0.50$ and $\mu_i^* = 0.25$, so $0 < \bar{\lambda}_i < \mu - \mu_j^*$ $(i=1,2)$. We find $\lambda_{11}^* = \lambda_{21}^* = 0.47$, $\lambda_{12}^* = \lambda_{22}^* = 0.17$, $\pi_{11}^* = \pi_{21}^* = 1.18$, $\pi_{12}^* = \pi_{22}^* = 0.43$.

4. The Cooperative Solution

In this section we suppose our two users agree to choose arrival rates λ_1 and λ_2 which maximize the sum of their profits. In effect then the two decision-makers are replaced by a single decision-maker with objective function

$$\psi(\lambda_1,\lambda_2;\mu) = \pi_1(\lambda_1;\mu,\lambda_2) + \pi_2(\lambda_2;\mu,\lambda_1). \tag{67}$$

Once more, the decision-maker will never choose $\lambda_1+\lambda_2 \geq \mu$ since in that case $\psi(\lambda_1,\lambda_2;\mu) = -\infty$. By (1),

$$\psi(\lambda_1,\lambda_2;\mu) = [R_1(\lambda_1)-h_1\lambda_1/(\mu-\lambda_1-\lambda_2)] + [R_2(\lambda_2)-h_2\lambda_2/(\mu-\lambda_1-\lambda_2)]$$

$$(\lambda_1+\lambda_2<\mu). \tag{68}$$

The question of how the maximized total profit, say ψ^o, is redistributed to the two users, we shall not discuss, except to state that it would seem reasonable that each user should receive as his share at least the equilibrium point profit, but no more than he might gain as a leader.

4.1. A Region Containing All Solutions

Let R_1, R_2, μ, h_1, h_2 be fixed and assume that Condition 2 holds for each constituent objective function, π_1 and π_2. Our first goal is the definition of a suitably small region which any $(\lambda_1,\lambda_2) = (\lambda_1^o,\lambda_2^o)$ maximizing ψ must lie in.

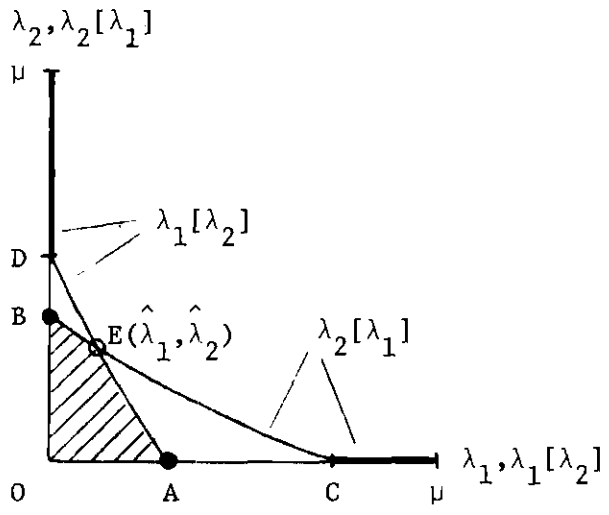

Figure IV-3. A Region Containing All Solutions.

Consider Figure IV-3 illustrating case (d). For the purpose of comparison we have indicated solutions under other behavioral assumptions dealt with in previous sections. The points $A=(\bar{\lambda}_1,0)$ and $B=(0,\bar{\lambda}_2)$ mark the one-user solutions; $E=(\hat{\lambda}_1,\hat{\lambda}_2)$ is the equilibrium point solution; the curve segments EC and ED, excluding E, contain the leader-follower solutions, namely $(\lambda^*_{11},\lambda^*_{22})$ and $(\lambda^*_{12},\lambda^*_{21})$ respectively. We will prove that a cooperative solution must lie in the set Y, shown as a shaded area, defined as the interior of the figure OBEA plus A and B.

For another illustration of the discussed solution concepts see Figure 8.10 in Intriligator [1971, p. 212].

Define

$$\psi^o = \sup_{\lambda_1,\lambda_2} \psi(\lambda_1,\lambda_2;\mu), \tag{69}$$

$$S = \{(\lambda^o_1,\lambda^o_2): \psi(\lambda^o_1,\lambda^o_2;\mu) = \psi^o\}. \tag{70}$$

S is the solution set for the cooperative model. A $(\lambda_1,\lambda_2) \in S$ will

usually be denoted by $(\lambda_1^o,\lambda_2^o)$.

Let the sets X, Y, Z be defined as

$$X = \{(\lambda_1,\lambda_2): 0 < \lambda_1 < \lambda_1[\lambda_2],\ 0 < \lambda_2 < \lambda_2[\lambda_1]\}, \tag{71}$$

$$Y = X \cup (\bar{\lambda}_1,0) \cup (0,\bar{\lambda}_2), \tag{72}$$

$$Z = \{(\lambda_1,\lambda_2): 0 \leq \lambda_1 \leq \lambda_1[\lambda_2],\ 0 \leq \lambda_2 \leq \lambda_2[\lambda_1]\}. \tag{73}$$

Clearly, $X \subset Y \subset Z$. In case (d), illustrated by Figure IV-3, X is the interior of the figure with corners O, B, E, A. Y includes also A and B, and Z is the closure of X. We will show that, in any case, $S \subset Y$.

First we show $S \subset Z$. Let $(\lambda_1,\lambda_2) \notin Z$. We must prove $(\lambda_1,\lambda_2) \notin S$. Suppose $\lambda_2 > \lambda_2[\lambda_1]$. If $\lambda_1 = 0$, then $\pi_1(\lambda_1;\mu,\lambda_2) = \pi_1(\lambda_1;\mu,\lambda_2[\lambda_1])$ and $\pi_2(\lambda_2;\mu,\lambda_1) < \pi_2(\lambda_2[\lambda_1];\mu,\lambda_1)$, so $\psi(\lambda_1,\lambda_2;\mu) < \psi(\lambda_1,\lambda_2[\lambda_1];\mu)$. If $\lambda_1 > 0$, then $\pi_1(\lambda_1;\mu,\lambda_2) < \pi_1(\lambda_1;\mu,\lambda_2[\lambda_1])$ and $\pi_2(\lambda_2;\mu,\lambda_1) < \pi_2(\lambda_2[\lambda_1];\mu,\lambda_1)$, so, again, $\psi(\lambda_1,\lambda_2;\mu) < \psi(\lambda_1,\lambda_2[\lambda_1];\mu)$. These results follow easily from Equation (1) and the fact that $\lambda_2[\lambda_1]$ uniquely maximizes the conditional objective function. Hence $(\lambda_1,\lambda_2) \notin S$. Similarly, if $\lambda_1 > \lambda_1[\lambda_2]$, then $(\lambda_1,\lambda_2) \notin S$. Since $(\lambda_1,\lambda_2) \notin Z$ means that either $\lambda_1 > \lambda_1[\lambda_2]$ or $\lambda_2 > \lambda_2[\lambda_1]$, we conclude $(\lambda_1,\lambda_2) \notin Z \Rightarrow (\lambda_1,\lambda_2) \notin S$. Equivalently, $S \subset Z$.

We have proved $S \subset Z$. We shall prove the stronger statement $S \subset Y$. Thus, all points on the boundary of Z, except $A=(\bar{\lambda}_1,0)$ and $B=(0,\bar{\lambda}_2)$ will be ruled out as elements of S.

First we rule out all remaining points on the axes, except A and B. Suppose $\lambda_1 = 0$ and $\lambda_2 < \lambda_2[0] = \bar{\lambda}_2$. Then $\pi_1(0;\mu,\lambda_2) = \pi_1(0;\mu,\lambda_2[0])$ and $\pi_2(\lambda_2;\mu,0) < \pi_2(\lambda_2[0];\mu,0)$, so $\psi(0,\lambda_2;\mu) < \psi(0,\lambda_2[0];\mu)$. Thus, if $\lambda_2 < \lambda_2[0]$ then $(0,\lambda_2) \notin S$. Similarly, if $\lambda_1 < \lambda_1[0]$ then $(\lambda_1,0) \notin S$.

Next we rule out all remaining points on the two curves $\lambda_1[\lambda_2]$ and $\lambda_2[\lambda_1]$, except A and B. Suppose $\lambda_1 > 0$ and $\lambda_2 = \lambda_2[\lambda_1] > 0$. Here we need to study the marginal change in ψ caused by a change in λ_2. By (67), (68), $d\psi/d\lambda_2 = d\pi_1/d\lambda_2 + d\pi_2/d\lambda_2$, where $d\pi_1/d\lambda_2 = -h_1\lambda_1/(\mu-\lambda_1-\lambda_2)^2 < 0$, and $d\pi_2/d\lambda_2 = 0$ by definition of $\lambda_2[\lambda_1]$ and the assumption that $\lambda_2 = \lambda_2[\lambda_1] > 0$. Hence, $d\psi/d\lambda_2 < 0$. It follows that a sufficiently small reduction in λ_2 will lead to an increase in ψ. Thus $(\lambda_1,\lambda_2[\lambda_1]) \notin S$. Similarly, if $\lambda_2 > 0$ and $\lambda_1 = \lambda_1[\lambda_2] > 0$, then $(\lambda_1[\lambda_2],\lambda_2) \notin S$. Our conclusion is that $S \subset Y$.

By our assumptions about π_1 and π_2, obviously ψ is continuous in λ_1 and λ_2 for $\lambda_1+\lambda_2 < \mu$. Hence, ψ is continuous on the compact set Z. By a well-known theorem it follows that ψ possesses a maximum value on Z. Thus S is nonempty.

In summary our findings are:

LEMMA 8. Let both users satisfy Condition 2. Let the objective function be $\psi(\lambda_1,\lambda_2;\mu) = \pi_1(\lambda_1;\mu,\lambda_2) + \pi_2(\lambda_2;\mu,\lambda_1)$. Then, $S \subset Y$, and S is nonempty.

The points $A = (\bar{\lambda}_1,0)$ and $B = (0,\bar{\lambda}_2)$ cannot, in general, be ruled out as solutions. Every case of linear reward functions, $R_i(\lambda_i) = r_i\lambda_i$ (i=1,2), provides a counterexample, as we shall see.

As stated in the lemma, a solution always exists. However, a solution is not necessarily unique as we shall see in the analysis of linear reward functions.

4.2. Solutions

In cases (a) and (b) there is the rather obvious unique solution

$(\lambda_1^o,\lambda_2^o) = (\hat{\lambda}_1,\hat{\lambda}_2)$. Thus we shall consider cases (c) and (d) only.

By Lemma 8, if neither $(\overline{\lambda}_1,0)$ nor $(0,\overline{\lambda}_2)$ maximizes ψ, then a solution will exist on the open set X. By our assumptions about π_1 and π_2, ψ is continuous and twice differentiable in λ_1 and λ_2 for $\lambda_1+\lambda_2< \mu$. Therefore, a necessary condition that a $(\lambda_1,\lambda_2) \in X$ maximizes ψ is that $d\psi/d\lambda_1 = 0$, $d\psi/d\lambda_2 = 0$. Hence, we will be looking for (λ_1,λ_2)'s satisfying

$$\frac{d\psi}{d\lambda_1} = \frac{d\pi_1}{d\lambda_1} + \frac{d\pi_2}{d\lambda_1} = \left[R_1'(\lambda_1) - \frac{h_1(\mu-\lambda_2)}{(\mu-\lambda_1-\lambda_2)^2} \right] - \frac{h_2\lambda_2}{(\mu-\lambda_1-\lambda_2)^2} = 0, \qquad (74)$$

$$\frac{d\psi}{d\lambda_2} = \frac{d\pi_2}{d\lambda_2} + \frac{d\pi_1}{d\lambda_2} = \left[R_2'(\lambda_2) - \frac{h_2(\mu-\lambda_1)}{(\mu-\lambda_1-\lambda_2)^2} \right] - \frac{h_1\lambda_1}{(\mu-\lambda_1-\lambda_2)^2} = 0. \qquad (75)$$

By Lemma 3, Condition 2 guarantees $d^2\pi_i/d\lambda_i^2 < 0$ (i=1,2) for all $\lambda_1+\lambda_2< \mu$, and evidently $d^2\pi_j/d\lambda_i^2 < 0$ for all $\lambda_1+\lambda_2< \mu$. Hence, $d^2\psi/d\lambda_i^2 < 0$ (i=1,2) for all $\lambda_1+\lambda_2< \mu$.

The economic interpretation of (74) and (75) is this: For user i (i=1,2) the marginal revenue per unit increase in arrival rate λ_i (i.e. $R_i'(\lambda_i)$) should equal the marginal waiting loss of the two users put together (i.e. $[h_i(\mu-\lambda_j)+h_j\lambda_j]/(\mu-\lambda_1-\lambda_2)^2$).

Simplification of (74), (75) results in the equivalent necessary conditions

$$R_1'(\lambda_1)-h_1/(\mu-\lambda_1-\lambda_2) = R_2'(\lambda_2)-h_2/(\mu-\lambda_1-\lambda_2) = (h_1\lambda_1+h_2\lambda_2)/(\mu-\lambda_1-\lambda_2)^2.$$

Interestingly, $R_i'(\lambda_i)-h_i/(\mu-\lambda_1-\lambda_2)$ is the marginal gain for π_i per unit increase in λ_i for fixed $\lambda_1+\lambda_2$.

In the special case $h_1 = h_2 = h$ (74), (75) reduce to

$$R_1'(\lambda_1) = R_2'(\lambda_2) = h\mu/(\mu-\lambda_1-\lambda_2)^2 \qquad (h_1=h_2=h). \tag{76}$$

Using (76) it will often be quite easy to locate all (λ_1,λ_2)'s satisfying the first-order necessary conditions for optimality, and thereby identify S.

4.3. Comparison with Equilibrium Solution

The definition of ψ, Lemma 8, and our results concerning the slope of $\lambda_1[\lambda_2]$ and $\lambda_2[\lambda_1]$, see Equation (32), lead to the following conclusions: In cases (a) and (b), $\lambda_1^o=\hat{\lambda}_1$, $\lambda_2^o=\hat{\lambda}_2$, $\psi^o=\hat{\pi}_1+\hat{\pi}_2$; in case (c), $\lambda_1^o+\lambda_2^o \leq \hat{\lambda}_1+\hat{\lambda}_2$, $\psi^o \geq \hat{\pi}_1+\hat{\pi}_2$; in case (d) $\lambda_1^o+\lambda_2^o < \hat{\lambda}_1+\hat{\lambda}_2$, $\psi^o > \hat{\pi}_1+\hat{\pi}_2$. We summarize our results as follows:

THEOREM 6. Let both users satisfy Condition 2. Suppose $(\lambda_1^o,\lambda_2^o) \in S$, i.e. $\psi(\lambda_1^o,\lambda_2^o;\mu)=\psi^o=\max_{\lambda_1,\lambda_2}\psi(\lambda_1,\lambda_2;\mu)$. Then $\lambda_1^o+\lambda_2^o \leq \hat{\lambda}_1+\hat{\lambda}_2$ and $\psi^o \geq \hat{\pi}_1+\hat{\pi}_2$, where inequality may hold.

4.4. Example. $R_1(\lambda_1) = r_1\lambda_1$, $R_2(\lambda_2) = r_2\lambda_2$

In this case ψ becomes

$$\psi(\lambda_1,\lambda_2;\mu) = r_1\lambda_1 + r_2\lambda_2 - (h_1\lambda_1+h_2\lambda_2)/(\mu-\lambda_1-\lambda_2), \tag{77}$$

which can be rewritten as

$$\begin{aligned}\psi(\lambda_1,\lambda_2;\mu) &= [r_2(\lambda_1+\lambda_2) - h_2(\lambda_1+\lambda_2)/(\mu-(\lambda_1+\lambda_2))] \\ &\quad + [(r_1-r_2) + (h_2-h_1)/(\mu-(\lambda_1+\lambda_2))]\cdot\lambda_1 \\ &= f(\lambda_1+\lambda_2) + g(\lambda_1+\lambda_2)\cdot\lambda_1,\end{aligned} \tag{78}$$

where the equation defines f and g as functions of $\lambda_1+\lambda_2$. Thus, for fixed

$\lambda_1+\lambda_2$, ψ is a linear function in λ_1 (or in λ_2). Simple results are a consequence of this property.

First of all, for any given $\lambda=\lambda_1+\lambda_2$ ψ will be maximized for $\lambda_1=0$ or $\lambda_1=\lambda$ depending on the sign of $g(\lambda_1+\lambda_2)$. Should $g(\lambda_1+\lambda_2)$ happen to be zero, then the choice of λ_1 does not matter. The linearity property thus implies that a search for a $(\lambda_1^o,\lambda_2^o)$ that maximizes ψ can be limited to the boundary set

$$L = \{(\lambda_1,0): \lambda_1 \geq 0\} \cup \{(0,\lambda_2): \lambda_2 \geq 0\}.$$

On the λ_1-axis the unique optimal solution is $(\overline{\lambda}_1,0)$, and on the λ_2-axis the unique optimal solution is $(0,\overline{\lambda}_2)$. Hence, either $(\overline{\lambda}_1,0)$ or $(0,\overline{\lambda}_2)$, or both, maximize ψ.

Suppose $r_1= r_2$, $h_1= h_2$. Then $\overline{\lambda}_1= \overline{\lambda}_2$ and $\psi^o = \psi(\overline{\lambda}_1,0;\mu) = \psi(0,\overline{\lambda}_2;\mu)$. Moreover, as $g(\lambda_1+\lambda_2) \equiv 0$, every point on the line segment connecting $(\overline{\lambda}_1,0)$ and $(0,\overline{\lambda}_2)$ also maximizes ψ, and no other point does so.

We will prove that an interior point (not in L) may maximize ψ only if $r_1= r_2$, $h_1= h_2$. Suppose a point $(\lambda_1^o,\lambda_2^o)$, with $\lambda_1> 0$ and $\lambda_2> 0$ maximizes ψ. Then $g(\lambda_1^o,\lambda_2^o) = 0$, since otherwise, by (78), ψ might be increased by a change in the arrival rates which leaves the total arrival rate at $\lambda_1^o+\lambda_2^o$. It follows that $(\lambda_1^o+\lambda_2^o,0)$ and $(0,\lambda_1^o+\lambda_2^o)$ both maximize ψ. By uniqueness of $\overline{\lambda}_1$ and $\overline{\lambda}_2$, we find $\overline{\lambda}_1= \overline{\lambda}_2= \lambda_1^o+\lambda_2^o$. Hence $\psi(\overline{\lambda}_1,0;\mu) = \psi(0,\overline{\lambda}_2;\mu) = \psi^o$. Now, by (28), $\overline{\lambda}_i= \mu - \sqrt{\mu h_i/r_i}$. Thus, $\overline{\lambda}_1=\overline{\lambda}_2$ implies $h_1/r_1 = h_2/r_2$. Alternatively, for some $k > 0$, $r_2= kr_1$ and $h_2= kh_1$. Inserting these expressions for r_2 and h_2 into $\psi(0,\overline{\lambda}_2;\mu)$ as given by

(77) we obtain

$$\psi(0,\bar{\lambda}_2;\mu) = r_2\bar{\lambda}_2 - h_2\bar{\lambda}_2/(\mu-\bar{\lambda}_2)$$

$$= k[r_1\bar{\lambda}_1 - h_1\bar{\lambda}_1/(\mu-\bar{\lambda}_1)] \qquad [\bar{\lambda}_1=\bar{\lambda}_2]$$

$$= k\ \psi(\bar{\lambda}_1,0;\mu).$$

We conclude that $k = 1$. That is, $r_1 = r_2$ and $h_1 = h_2$, as was to be shown.

Suppose $\bar{\lambda}_1 > 0$, $\bar{\lambda}_2 > 0$. Then both $A = (\bar{\lambda}_1,0)$ and $B = (0,\bar{\lambda}_2)$ may maximize ψ even if $r_1 \neq r_2$ or $h_1 \neq h_2$. Parameter sets with this property are easily constructed. Take an arbitrary parameter set $\mu^{(1)}$, $r_1^{(1)}$, $h_1^{(1)}$, $r_2^{(1)}$, $h_2^{(1)}$ such that $h_1^{(1)}/r_1^{(1)} < \mu^{(1)}$, $h_2^{(1)}/r_2^{(1)} < \mu^{(1)}$ and $h_1^{(1)}/r_1^{(1)} \neq h_2^{(1)}/r_2^{(1)}$. By results in Section 1, $\bar{\lambda}_1^{(1)} > 0$, $\bar{\lambda}_2^{(1)} > 0$, $\bar{\lambda}_1^{(1)} \neq \bar{\lambda}_2^{(1)}$. There exists $k > 0$ such that $\psi^{(1)}(\bar{\lambda}_1^{(1)},0;\mu^{(1)}) = k\psi^{(1)}(0,\bar{\lambda}_2^{(1)};\mu^{(1)})$. Now consider the parameter set $\mu^{(2)} = \mu^{(1)}$, $r_1^{(2)} = r_1^{(1)}$, $h_1^{(2)} = h_1^{(1)}$, $r_2^{(2)} = kr_2^{(1)}$, $h_2^{(2)} = kh_2^{(1)}$. Clearly, $\bar{\lambda}_1^{(2)} = \bar{\lambda}_1^{(1)}$, $\bar{\lambda}_2^{(2)} = \bar{\lambda}_2^{(1)}$, and $\psi^{(2)}(\bar{\lambda}_1^{(2)},0;\mu^{(2)}) = \psi^{(1)}(\bar{\lambda}_1^{(1)},0;\mu^{(1)})$. By (77) $\psi^{(2)}(0,\bar{\lambda}_2^{(2)};\mu^{(2)}) = k\psi^{(1)}(0,\bar{\lambda}_2^{(1)};\mu^{(1)})$. Hence, $\psi^{(2)}(\bar{\lambda}_1^{(2)},0;\mu^{(2)}) = \psi^{(2)}(0,\bar{\lambda}_2^{(2)};\mu^{(2)})$, proving our assertion.

In particular cases, $g(\lambda_1+\lambda_2)$ is either positive or negative for all $\lambda=\lambda_1+\lambda_2$, and it is easy to determine whether $(\bar{\lambda}_1,0)$ or $(0,\bar{\lambda}_2)$ is the solution. Specifically, if $r_1 \geq r_2$ and $h_1 < h_2$, or, $r_1 > r_2$ and $h_1 = h_2$, then $S = (\bar{\lambda}_1,0)$; and if $r_1 \leq r_2$ and $h_1 > h_2$, or, $r_1 < r_2$ and $h_1 = h_2$, then $S = (0,\bar{\lambda}_2)$.

We summarize our findings as follows.

THEOREM 7. Let $\psi(\lambda_1,\lambda_2;\mu)$ be given by Equation (77). Then,

(i) If $r_1 = r_2$ and $h_1 = h_2$, then $\overline{\lambda}_1 = \overline{\lambda}_2$ and
$S = \{(\lambda_1,\lambda_2): (\lambda_1,\lambda_2) = (\alpha\overline{\lambda}_1,(1-\alpha)\overline{\lambda}_1),\ 0 \leq \alpha \leq 1\}$,

(ii) If $r_1 \neq r_2$ or $h_1 \neq h_2$, then S equals $(\overline{\lambda}_1,0)$, or $(0,\overline{\lambda}_2)$, or $(\overline{\lambda}_1,0) \cup (0,\overline{\lambda}_2)$.

Finally, let $\psi_{11} = d^2\psi/d\lambda_1^2$, $\psi_{22} = d^2\psi/d\lambda_2^2$, $\psi_{12} = \psi_{21} = d^2\psi/d\lambda_1 d\lambda_2$. Let Δ denote the 2x2 Hessian determinant. We derive

$$\Delta = \begin{vmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{vmatrix} = -\frac{(h_1-h_2)^2}{(\mu-\lambda_1-\lambda_2)^4} \leq 0 \qquad (\lambda_1+\lambda_2<\mu). \tag{79}$$

A nonpositive Δ implies that nowhere on the interior $(\lambda_1>0,\lambda_2>0)$ can ψ have a (strict) local maximum or minimum. This agrees with our results, since $\Delta \leq 0$ does not rule out a local or global maximum on the boundary L; nor does it rule out a local nonstrict maximum at points in the interior.

In our numerical example with $\mu = 1$, $r_1 = r_2 = 8$, $h_1 = h_2 = 2$, we have $\overline{\lambda}_1 = \overline{\lambda}_2 = 0.50$. Hence, $\psi^o = \psi(\overline{\lambda}_1,0;\mu) = 2.00$ and the solution set is $S = \{(\lambda_1,\lambda_2): (\lambda_1,\lambda_2) = (\alpha 0.50,(1-\alpha)0.50),\ 0 \leq \alpha \leq 1\}$.

5. Appendix

In this section we give proof of Lemmas 1 through 5 of Section 1. To avoid confusion we usually will denote by $\overline{\lambda}(\mu)$ the optimal arrival rate at service rate μ rather than by $\overline{\lambda}$. Similarly we use the notation $\overline{\phi}(\mu) \equiv \phi(\overline{\lambda}(\mu);\mu)$ instead of $\overline{\phi}$. When h is treated as a variable, as in Lemma 2, its value will be indicated by the usual "conditional upon"

or "given" notation.

5.1. Proof of Lemma 1

Let $\Lambda_\infty = \{\lambda: 0<\lambda<\infty, R(\lambda)>R(0)\}$. By assumption $\Lambda_\infty \neq \emptyset$. Obviously, $\overline{\lambda}(\mu) \in \{0\} \cup \Lambda_\infty$, for any μ. We shall derive the set of μ's such that $\overline{\lambda}(\mu) = 0$, for given R and h.

For fixed $\lambda \in \Lambda_\infty$ consider $\phi(\lambda;\mu)$ as a function of μ. By (7), $\phi(\lambda;\mu) = -\infty$ for $\mu \leq \lambda$, and on the interval (λ,∞) $\phi(\lambda;\mu)$ is continuous and strictly increasing in μ, with $\lim_{\mu\to\lambda} \phi(\lambda;\mu) = -\infty$ and $\lim_{\mu\to\infty} \phi(\lambda;\mu) = R(\lambda) > R(0)$. By the intermediate value theorem it follows that there is a (unique) μ, say $\mu(\lambda) > \lambda$, such that $\phi(\lambda;\mu(\lambda)) = R(0)$.

Let $\mu^* = \inf_{\lambda\in\Lambda_\infty} \mu(\lambda)$. We will show $\mu^* > 0$. Recall that, by assumption, $R'(\lambda) \leq M'$ for all λ, with $0 < M' < \infty$. Hence, for $0 < \lambda < \mu$,

$$\begin{aligned}\phi(\lambda;\mu) &= R(\lambda) - h\lambda/(\mu-\lambda) \\ &\leq R(0) + \lambda[M' - h/(\mu-\lambda)] \\ &< R(0) + \lambda[M' - h/\mu].\end{aligned}$$

It is seen that $\phi(\lambda;\mu) < R(0)$ for $0 < \lambda < \mu \leq h/M'$. Also $\phi(\lambda;\mu) = -\infty < R(0)$ if $\lambda > 0$ and $\mu \leq \lambda$. Hence,

$$\phi(\lambda;\mu) < R(0) \qquad (\lambda>0;\ \mu\leq h/M').$$

In particular, $\phi(\lambda;\mu) < R(0)$ for all $\lambda \in \Lambda_\infty$ and all $\mu \leq h/M'$. Therefore, $\mu(\lambda) > h/M'$ for all $\lambda \in \Lambda_\infty$. Hence, $\mu^* \geq h/M' > 0$.

By the definition of μ^*, clearly $\overline{\lambda}(\mu) = 0$ for $\mu < \mu^*$, and $\overline{\lambda}(\mu) > 0$ for $\mu > \mu^*$. The question then is, what is $\overline{\lambda}(\mu^*)$? If there is no $\lambda_o \in \Lambda_\infty$ such that $\mu(\lambda_o) = \mu^*$, then $\phi(\lambda;\mu^*) < \phi(0;\mu^*) = R(0)$ for all $\lambda \in \Lambda_\infty$, so $\overline{\phi}(\mu^*) = R(0)$ and $\overline{\lambda}(\mu^*) = 0$. If, on the other hand, there is

a $\lambda_o \varepsilon \Lambda_\infty$ such that $\mu(\lambda_o) = \mu^*$, then $\phi(\lambda_o;\mu^*) = \phi(0;\mu^*) = \overline{\phi}(\mu^*) = R(0)$. But, by assumption, we always choose the smallest λ satisfying $\phi(\lambda;\mu) = \overline{\phi}(\mu)$, so again $\overline{\phi}(\mu^*) = R(0)$ and $\overline{\lambda}(\mu^*) = 0$. We have shown that $\overline{\lambda}(\mu) = 0$ if and only if $\mu \leq \mu^*$, for the given R and h.

Next, we shall show that $\overline{\phi}(\mu)$ and $\overline{\lambda}(\mu)$ are strictly increasing in μ on $[\mu^*,\infty)$. By the above analysis, for $\mu > \mu^*$ we have $\overline{\phi}(\mu) > \overline{\phi}(\mu^*) = R(0)$ and $\overline{\lambda}(\mu) > \overline{\lambda}(\mu^*) = 0$, so we need only show that $\overline{\phi}$ and $\overline{\lambda}$ are strictly increasing in μ on (μ^*,∞). Suppose $\mu^* < \mu_a < \mu_b$. We must prove $\overline{\phi}(\mu_a) < \overline{\phi}(\mu_b)$, $\overline{\lambda}(\mu_a) < \overline{\lambda}(\mu_b)$.

Note, $0 < \overline{\lambda}(\mu_a) < \mu_a$, $0 < \overline{\lambda}(\mu_b) < \mu_b$. By (7) and the definition of $\overline{\lambda}$,

$$\overline{\phi}(\mu_a) \equiv \phi(\overline{\lambda}(\mu_a);\mu_a) < \phi(\overline{\lambda}(\mu_a);\mu_b) \leq \phi(\overline{\lambda}(\mu_b);\mu_b) \equiv \overline{\phi}(\mu_b).$$

Hence, $\overline{\phi}(\mu_a) < \overline{\phi}(\mu_b)$. Thus $\overline{\phi}$ is strictly increasing on (μ^*,∞), and therefore on $[\mu^*,\infty)$.

The arguments needed for proof of $\overline{\lambda}(\mu_a) < \overline{\lambda}(\mu_b)$ are more involved. We begin by showing that $\overline{\lambda}(\mu_a) \leq \overline{\lambda}(\mu_b)$. Our starting point is the inequality

$$\phi(\overline{\lambda}(\mu_a);\mu_a) > \phi(\lambda;\mu_a) \qquad (\lambda<\overline{\lambda}(\mu_a)),$$

implied by the definition of $\overline{\lambda}(\mu_a)$. Now, by (11), for given $\lambda > 0$, the waiting loss $z = h\lambda/(\mu-\lambda)$ is a decreasing function of μ on (λ,∞), and by (12) the rate of decrease is greater the greater λ is. Hence, the rate of increase in ϕ is greater at $\overline{\lambda}(\mu_a)$ than at any $\lambda < \overline{\lambda}(\mu_a)$. Consequently,

$$\phi(\overline{\lambda}(\mu_a);\mu_b) - \phi(\overline{\lambda}(\mu_a);\mu_a) > \phi(\lambda;\mu_b) - \phi(\lambda;\mu_a) \qquad (\lambda<\overline{\lambda}(\mu_a)).$$

"Adding" the two above inequalities we obtain

$$\phi(\overline{\lambda}(\mu_a);\mu_b) > \phi(\lambda;\mu_b) \qquad (\lambda<\overline{\lambda}(\mu_a)).$$

Evidently, $\phi(\overline{\lambda}(\mu_b);\mu_b) \geq \phi(\overline{\lambda}(\mu_a);\mu_b)$. Hence

$$\phi(\overline{\lambda}(\mu_b);\mu_b) > \phi(\lambda;\mu_b) \qquad (\lambda<\overline{\lambda}(\mu_a)).$$

From this we deduce that $\overline{\lambda}(\mu_a) \leq \overline{\lambda}(\mu_b)$.

We can also rule out the possibility $\overline{\lambda}(\mu_a) = \overline{\lambda}(\mu_b)$ as

$$\left.\frac{d\phi(\lambda;\mu_b)}{d\lambda}\right|_{\overline{\lambda}(\mu_a)} = R'(\overline{\lambda}(\mu_a))-h\mu_b/(\mu_b-\overline{\lambda}(\mu_a))^2$$

$$> R'(\overline{\lambda}(\mu_a))-h\mu_a/(\mu_a-\overline{\lambda}(\mu_a))^2 \qquad [\overline{\lambda}(\mu_a)<\mu_a<\mu_b]$$

$$= 0 \qquad [\overline{\lambda}(\mu_a)>0 \Rightarrow \phi'(\overline{\lambda}(\mu_a);\mu_a)=0].$$

(To see that the inequality holds write $\mu/(\mu-\lambda)^2$ as $(\mu-\lambda)^{-1}(1-\lambda/\mu)^{-1}$.) We conclude that $\overline{\lambda}(\mu_a) < \overline{\lambda}(\mu_b)$. Thus $\overline{\lambda}$ is strictly increasing on (μ^*,∞) and therefore on $[\mu^*,\infty)$.

It remains to discuss the question of continuity. First consider $\overline{\phi}$. We shall prove that $\overline{\phi}$ is continuous in μ, for any given R and h. $\overline{\phi}(\mu) = R(0)$ for $\mu \leq \mu^*$, so clearly $\overline{\phi}$ is continuous on $(0,\mu^*]$. We must prove $\overline{\phi}$ is continuous on $[\mu^*,\infty)$. Let $\mu^* \leq \mu_a < \mu_b < \infty$. Our first step is to show that for a fixed μ_a (μ_b) a sufficiently close value of μ_b (μ_a) will ensure $\overline{\lambda}(\mu_b) < \mu_a$.

Fix $\mu_b > \mu^*$. As always, $\overline{\lambda}(\mu_b) < \mu_b$. Obviously, any $\mu_a \in (\max[\mu^*,\overline{\lambda}(\mu_b)],\mu_b)$ will satisfy $\overline{\lambda}(\mu_b) < \mu_a$.

Fix $\mu_a \geq \mu^*$. For any $\mu_b > \mu_a$, clearly $0 \leq \overline{\lambda}(\mu_a) < \overline{\lambda}(\mu_b)$. Thus, by

(10), $\overline{\lambda}(\mu_b) \leq \mu_b - \sqrt{\mu_b h/M'}$. Therefore, if μ_b is chosen so that $\mu_b < \mu_a + \sqrt{\mu_b h/M'}$, then $\overline{\lambda}(\mu_b) < \mu_a$ as desired. Hence, if $\mu_b < \mu_a + \sqrt{\mu_a h/M'}$, then $\overline{\lambda}(\mu_b) < \mu_a$. We have shown that if, for fixed μ_a or μ_b, μ_a and μ_b are sufficiently close, then $\overline{\lambda}(\mu_b) < \mu_a$.

Suppose $\mu_b > \mu^*$ is fixed and $\mu_a \in (\max[\mu^*, \overline{\lambda}(\mu_b)], \mu_b)$ is variable; or suppose $\mu_a \geq \mu^*$ is fixed and $\mu_b \in (\mu_a, \mu_a + \sqrt{\mu_a h/M'})$ is variable. In either case $\overline{\lambda}(\mu_b) < \mu_a$ and

$$0 < \overline{\phi}(\mu_b) - \overline{\phi}(\mu_a) \qquad [\overline{\phi} \text{ is strictly increasing on } [\mu^*, \infty)]$$

$$\equiv \phi(\overline{\lambda}(\mu_b); \mu_b) - \phi(\overline{\lambda}(\mu_a); \mu_a)$$

$$\leq \phi(\overline{\lambda}(\mu_b); \mu_b) - \phi(\overline{\lambda}(\mu_b); \mu_a) \qquad [\text{by definition of } \overline{\lambda}]$$

$$= \frac{h\overline{\lambda}(\mu_b)}{\mu_a - \overline{\lambda}(\mu_b)} - \frac{h\overline{\lambda}(\mu_b)}{\mu_b - \overline{\lambda}(\mu_b)} \qquad [\overline{\lambda}(\mu_b) < \mu_a; \text{ Eq.(7)}].$$

Taking limits one obtains

$$0 \leq \lim_{\mu_a \to \mu_b} (\overline{\phi}(\mu_b) - \overline{\phi}(\mu_a)) \leq 0 \quad (\mu_b > \mu^*;\ \mu_a \in (\max[\mu^*, \overline{\lambda}(\mu_b)], \mu_b)),$$

$$0 \leq \lim_{\mu_b \to \mu_a} (\overline{\phi}(\mu_b) - \overline{\phi}(\mu_a)) \leq 0 \quad (\mu_a \geq \mu^*;\ \mu_b \in (\mu_a, \mu_a + \sqrt{\mu_a h/M'})).$$

Hence $\lim_{\mu_a \to \mu_b} \overline{\phi}(\mu_a) = \overline{\phi}(\mu_b)$, and $\lim_{\mu_b \to \mu_a} \overline{\phi}(\mu_b) = \overline{\phi}(\mu_a)$. This proves that $\overline{\phi}$ is continuous on $[\mu^*, \infty)$. Since $\overline{\phi}$ is also continuous on $(0, \mu^*]$, we conclude that $\overline{\phi}$ is continuous in μ on $(0, \infty)$, for any given R and h.

Finally, we show by counterexample that $\overline{\lambda}$ is not, in general, continuous in μ. Take the case of a reward function R with two local maxima, one at λ_a, another at λ_b, where $0 < \lambda_a < \lambda_b$ and $R(0) < R(\lambda_a) < R(\lambda_b)$. For sufficiently small μ, clearly $\overline{\lambda}(\mu) < \lambda_a$. For sufficiently

large μ, clearly $\lambda_b - \varepsilon < \overline{\lambda}(\mu) < \lambda_b$ for arbitrary $\varepsilon > 0$. Equally clear, for no μ will $\overline{\lambda}(\mu)$ be such that $R'(\overline{\lambda}) \leq 0$. Thus, as μ increases, $\overline{\lambda}$ will jump from less than λ_a to some $\lambda > \lambda_a$ such that $R'(\lambda) > 0$. □

5.2. Proof of Lemma 2

Let $\Lambda_\mu = \{\lambda: 0<\lambda<\mu, R(\lambda)>R(0)\}$. Consider first the trivial case $\Lambda_\mu = \emptyset$. Evidently $\overline{\lambda}(\mu)\big|_h = 0$ for all $h \geq h^* = 0$. It follows that in the present case $\overline{\lambda}(\mu)\big|_h = 0$ if and only if $h \geq h^* = 0$.

Now consider the case $\Lambda_\mu \neq \emptyset$. For fixed $\lambda \in \Lambda_\mu$, let us examine $\phi(\lambda;\mu)\big|_h$ as a function of h. By (7) ϕ is strictly decreasing in h. The solution of $\phi(\lambda;\mu)\big|_h = R(0)$ is easily found to be $h(\lambda) = [R(\lambda)-R(0)]\cdot[(\mu/\lambda)-1] > 0$. Since, for every $\lambda \in \Lambda_\mu$, $h(\lambda) < \mu[R(\lambda)-R(0)]/\lambda \leq \mu M'$, it is clear there is an upper bound to $h(\lambda)$. Let $h^* = \sup_{\lambda\in\Lambda_\mu} h(\lambda)$ denote the least upper bound. Clearly, $h^* > 0$ as $h(\lambda) > 0$ for all $\lambda \in \Lambda_\mu$. The definition of h^* implies that $\overline{\lambda}(\mu)\big|_h = 0$ if $h > h^*$ and $\overline{\lambda}(\mu)\big|_h > 0$ if $h < h^*$. By arguments similar to those employed in the proof of $\overline{\lambda}(\mu^*) = 0$ in Lemma 1 it can be shown that $\overline{\lambda}(\mu)\big|_{h^*} = 0$. Hence, if $\Lambda_\mu \neq \emptyset$ there exists an h^*, $0 < h^* < \infty$, such that $\overline{\lambda}(\mu)\big|_h = 0$ if and only if $h \geq h^*$.

Next we will show that $\overline{\phi}$ and $\overline{\lambda}$ are strictly decreasing in h on $(0,h^*]$. By the above, $\overline{\lambda}(\mu)\big|_h > \overline{\lambda}(\mu)\big|_{h^*} = 0$ and $\overline{\phi}(\mu)\big|_h > \overline{\phi}(\mu)\big|_{h^*} = R(0)$ for all $h < h^*$. Thus we need to show only that $\overline{\phi}$ and $\overline{\lambda}$ are strictly decreasing in h on $(0,h^*)$. Assume therefore $0 < h_a < h_b < h^*$. For notational convenience, let $\overline{\lambda}_a = \overline{\lambda}(\mu)\big|_{h_a}$, $\overline{\lambda}_b = \overline{\lambda}(\mu)\big|_{h_b}$. Note, $\overline{\lambda}_a > 0$ and $\overline{\lambda}_b > 0$.

First we prove that $\overline{\phi}$ is strictly decreasing in h on $(0,h^*)$. By (7) and the definition of $\overline{\lambda}$

$$\overline{\phi}(\mu)\big|_{h_a} \equiv \phi(\overline{\lambda}_a;\mu)\big|_{h_a} \geq \phi(\overline{\lambda}_b;\mu)\big|_{h_a} > \phi(\overline{\lambda}_b;\mu)\big|_{h_b} \equiv \overline{\phi}(\mu)\big|_{h_b}.$$

Thus $\bar{\phi}(\mu)\big|_{h_a} > \bar{\phi}(\mu)\big|_{h_b}$, proving our statement.

Second we prove that $\bar{\lambda}$ is strictly decreasing in h on $(0,h^*)$. Proceeding along the lines of the proof of Lemma 1 we show successively

$$\phi(\bar{\lambda}_b;\mu)\big|_{h_b} > \phi(\lambda;\mu)\big|_{h_b} \qquad (\lambda<\bar{\lambda}_b),$$

$$\phi(\bar{\lambda}_b;\mu)\big|_{h_a} - \phi(\bar{\lambda}_b;\mu)\big|_{h_b} > \phi(\lambda;\mu)\big|_{h_a} - \phi(\lambda;\mu)\big|_{h_b} \qquad (\lambda<\bar{\lambda}_b),$$

$$\phi(\bar{\lambda}_b;\mu)\big|_{h_a} > \phi(\lambda;\mu)\big|_{h_a} \qquad (\lambda<\bar{\lambda}_b),$$

$$\phi(\bar{\lambda}_a;\mu)\big|_{h_a} > \phi(\lambda;\mu)\big|_{h_a} \qquad (\lambda<\bar{\lambda}_b).$$

It follows that $\bar{\lambda}_a \geq \bar{\lambda}_b$. The possibility $\bar{\lambda}_a = \bar{\lambda}_b$ can also be ruled out as

$$\left.\frac{d\phi(\lambda;\mu)}{d\lambda}\right|_{h_b,\bar{\lambda}_a} = R'(\bar{\lambda}_a) - h_b\mu/(\mu-\bar{\lambda}_a)^2$$

$$< R'(\bar{\lambda}_a) - h_a\mu/(\mu-\bar{\lambda}_a)^2$$

$$= 0 \qquad [\bar{\lambda}_a > 0 \Rightarrow \phi'(\bar{\lambda}_a;\mu)\big|_{h_a} = 0].$$

We conclude that $\bar{\lambda}_a > \bar{\lambda}_b$. Thus $\bar{\lambda}$ is strictly decreasing on $(0,h^*)$ and therefore on $(0,h^*]$.

Obviously, $\bar{\phi}$ is continuous in h on $[h^*,\infty)$ since $\bar{\phi} = R(0)$ on that interval. We shall prove continuity on $(0,h^*]$. Let $0 < h_a < h_b \leq h^*$. Then

$$0 < \bar{\phi}(\mu)\big|_{h_a} - \bar{\phi}(\mu)\big|_{h_b} \qquad [\bar{\phi} \text{ is strictly decreasing on } (0,h^*]]$$

$$\equiv \phi(\bar{\lambda}_a;\mu)\big|_{h_a} - \phi(\bar{\lambda}_b;\mu)\big|_{h_b}$$

$$\leq \phi(\bar{\lambda}_a;\mu)\big|_{h_a} - \phi(\bar{\lambda}_a;\mu)\big|_{h_b} \qquad [\text{by definition of } \bar{\lambda}]$$

$$= \frac{h_b \bar{\lambda}_a}{\mu - \bar{\lambda}_a} - \frac{h_a \bar{\lambda}_a}{\mu - \bar{\lambda}_a} \qquad \text{[by Eq.(7)]}$$

$$= (h_b - h_a)\bar{\lambda}_a / (\mu - \bar{\lambda}_a).$$

Whether $h_a \to h_b$ or $h_b \to h_a$ we derive easily $\lim (\bar{\phi}(\mu)|_{h_a} - \bar{\phi}(\mu)|_{h_b}) = 0$. This proves that $\bar{\phi}$ is continuous on $(0,h^*]$. Since $\bar{\phi}$ is also continuous on $[h^*,\infty)$ we conclude that $\bar{\phi}$ is continuous in h on $(0,\infty)$, for any given R and μ.

Finally, the counterexample at the end of Section 5.1 (add that $\lambda_b < \mu$) also serves to prove that $\bar{\lambda}$ is not, in general, continuous in h. There is no need for repeating the argument. □

5.3. Proof of Lemma 3

We begin by proving the statements about the values of μ which satisfy (17) (Condition 1). Clearly, if $R''(\lambda) \leq 0$ for all $\lambda \geq 0$, then (17) holds for all μ, and if we set $\mu^{(1)} = \infty$ then it is evident that $\mu < \mu^{(1)} \Leftrightarrow$ (17).

Now consider the alternative, $R''(\lambda) > 0$ for some $\lambda \geq 0$. As a first step we demonstrate that (17) is satisfied for sufficiently small μ. Toward this end, take any $\lambda_o > 0$. By assumption, R'' is continuous, so R'' will have a maximum $R_o'' < \infty$ somewhere on the closed interval $[0,\lambda_o]$. If $R_o'' \leq 0$ then it is clear that (17) is satisfied for all $\mu < \lambda_o$. Thus suppose $R_o'' > 0$. Now, $2h\mu/(\mu-\lambda)^3$ is strictly increasing in λ on $[0,\mu)$ and attains its minimum value $2h/\mu^2$ for $\lambda = 0$. Solving $R_o'' = 2h/\mu^2$ for μ we get $\bar{\mu} = \sqrt{2h/R_o''}$. Now let $\mu_o = \min(\bar{\mu},\lambda_o) > 0$. We see that $\mu < \mu_o \Rightarrow$ (17).

For each $\lambda \geq 0$, let $\mu^{(1)}(\lambda) = \infty$ if $R''(\lambda) \leq 0$. If, on the other

hand, $R''(\lambda) > 0$, then define $\mu^{(1)}(\lambda)$, $\lambda < \mu^{(1)}(\lambda) < \infty$, as the solution of $R''(\lambda) = 2h\mu/(\mu-\lambda)^3$ with respect to μ. Now, $2h\mu/(\mu-\lambda)^3$ is strictly decreasing in μ on (λ,∞). Hence, in any case,

$$\lambda < \mu < \mu^{(1)}(\lambda) \quad \Leftrightarrow \quad R''(\lambda) < 2h\mu/(\mu-\lambda)^3 \qquad (\lambda \geq 0).$$

Let $\mu^{(1)} \equiv \inf_\lambda \mu^{(1)}(\lambda) < \infty$. From the above relation we deduce that (17) is satisfied for $\mu < \mu^{(1)}$, but is not satisfied for $\mu \geq \mu^{(1)}$. We know that (17) is satisfied for sufficiently small μ. Our conclusion is that there exists a $\mu^{(1)}$, $0 < \mu^{(1)} < \infty$, such that $\mu < \mu^{(1)} \Leftrightarrow$ (17), with $\mu^{(1)}$ as defined.

The statements concerning the values of μ which satisfy (18) (Condition 2) are proved in the same way, so we omit the proof.

It remains to show that if $R''(\lambda) > 0$ for some $\lambda \geq 0$, then $\mu^{(1)} > \mu^{(2)}$. We have $2h\mu/(\mu-\lambda)^3 > h/(\mu-\lambda)^2$ for $\lambda < \mu$. It follows that if $R''(\lambda) > 0$ then $\mu^{(1)}(\lambda) > \mu^{(2)}(\lambda)$, where $\mu^{(2)}(\lambda) > \lambda$ is the solution of $R''(\lambda) = h/(\mu-\lambda)^2$ for μ. Hence $\mu^{(1)} \geq \mu^{(2)}$. We can rule out $\mu^{(1)} = \mu^{(2)}$, so $\mu^{(1)} > \mu^{(2)}$. □

5.4. Proof of Lemma 4

We first prove the statements concerning values of h which satisfy (17) (Condition 1). Clearly, if $R''(\lambda) \leq 0$ for all $\lambda < \mu$, then (17) holds for all h, and if we set $h^{(1)} = 0$, then $h > h^{(1)} \Leftrightarrow$ (17).

Now consider the case where $R''(\lambda) > 0$ for some $\lambda \in [0,\mu)$. We will show that (17) is satisfied for sufficiently large h. By continuity, R'' possesses a maximum on $[0,\mu]$. Let $R_o'' = \max_{\lambda \leq \mu} R''(\lambda)$, where $0 < R_o'' < \infty$. Denote by h_o the solution of $R_o'' = 2h/\mu^2$. Thus $h_o = R_o''\mu^2/2 > 0$. Clearly,

$h > h_o \Leftrightarrow (17)$.

For each $\lambda \in [0,\mu)$ let $h^{(1)}(\mu) = 0$ if $R''(\lambda) \leq 0$. If, however, $R''(\lambda) > 0$, then $h^{(1)}(\lambda)$ is defined as the solution of $R''(\lambda) = 2h\mu/(\mu-\lambda)^3$. Thus $h^{(1)}(\lambda) = R''(\lambda)(\mu-\lambda)^3/(2\mu)$. As $2h\mu/(\mu-\lambda)^3$ is strictly increasing in h

$$h > h^{(1)}(\lambda) \quad \Leftrightarrow \quad R''(\lambda) < 2h\mu/(\mu-\lambda)^3 \qquad (0 \leq \lambda < \mu).$$

Let $h^{(1)} = \sup_\lambda h^{(1)}(\lambda) > 0$. From the above relation we deduce that (17) is satisfied for $h > h^{(1)}$, but is not satisfied for $h \leq h^{(1)}$. We know that (17) is satisfied for sufficiently large h, so we conclude there exists an $h^{(1)}$, $0 < h^{(1)} < \infty$, such that $h > h^{(1)} \Leftrightarrow (17)$, with $h^{(1)}$ as defined.

The statements concerning the values of h which satisfy (18) (Condition 2) are proved in the same way, so we omit the proof.

Finally, we must show that if $R''(\lambda) > 0$ for some $\lambda \in [0,\mu)$, then $h^{(1)} < h^{(2)}$. Let $h^{(2)}(\lambda)$ be the solution of $R''(\lambda) = h/(\mu-\lambda)^2$ for h. Thus $h^{(2)}(\lambda) = R''(\lambda)(\mu-\lambda)^2$. It is seen that $h^{(1)}(\lambda) < \frac{1}{2} h^{(2)}(\lambda)$ for all $\lambda < \mu$. Hence, $h^{(1)} \leq \frac{1}{2} h^{(2)} < h^{(2)}$. □

5.5. Proof of Lemma 5

The two parts of the lemma are proved in almost identical manner. Therefore we give only the proof of the first part.

Let R and h be given. We shall prove that $\overline{\lambda}$ is continuous in μ on $(0,\mu^{(1)})$. If $\mu^{(1)} \leq \mu^*$, then $\overline{\lambda}(\mu) = 0$ for all $\mu < \mu^{(1)}$, so in this case clearly $\overline{\lambda}$ is continuous in μ on $(0,\mu^{(1)})$. Assume therefore $\mu^{(1)} > \mu^*$. As $\overline{\lambda}(\mu) = 0$ for $\mu \leq \mu^*$, $\overline{\lambda}$ is continuous in μ on $(0,\mu^*]$. Thus, we have to

prove continuity on $[\mu^*,\mu^{(1)})$.

We begin by proving continuity on the open interval $(\mu^*,\mu^{(1)})$. Consider any $\mu_o \in (\mu^*,\mu^{(1)})$. The associated optimal value of λ is $\overline{\lambda}(\mu_o)$. Choose an $\varepsilon > 0$ small enough so that $\overline{\lambda}(\mu_o)-\varepsilon > 0$ and $\overline{\lambda}(\mu_o)+\varepsilon < \mu_o$. We must show there exists a $\delta > 0$ such that $|\mu-\mu_o| < \delta \Rightarrow |\overline{\lambda}(\mu)-\overline{\lambda}(\mu_o)| < \varepsilon$.

By Lemma 3, Condition 1 is met for $\mu=\mu_o < \mu^{(1)}$. Thus $\phi(\lambda;\mu_o)$ is strictly concave in λ on $[0,\mu_o)$. Hence,

$$\phi(\overline{\lambda}(\mu_o);\mu_o) > \phi(\overline{\lambda}(\mu_o)-\varepsilon;\mu_o),$$

$$\phi(\overline{\lambda}(\mu_o);\mu_o) > \phi(\overline{\lambda}(\mu_o)+\varepsilon;\mu_o).$$

Now, $\phi(\lambda;\mu)$ is continuous in μ for any fixed $\lambda < \mu$. It follows that there is a $\delta > 0$ such that

$$\phi(\overline{\lambda}(\mu_o);\mu) > \phi(\overline{\lambda}(\mu_o)-\varepsilon;\mu),$$

$$\phi(\overline{\lambda}(\mu_o);\mu) > \phi(\overline{\lambda}(\mu_o)+\varepsilon;\mu),$$

for all $\mu \in (\mu_o-\delta,\mu_o+\delta) \subset (\mu^*,\mu^{(1)})$. Moreover, for each such μ, $\mu < \mu^{(1)}$, so, by Lemma 3, Condition 1 is met for each $\mu \in (\mu_o-\delta,\mu_o+\delta)$. Hence $\phi(\lambda;\mu)$ is strictly concave in λ on $[0,\mu)$. By the two above inequalities, $\phi(\lambda;\mu)$ will have its maximum for a $\lambda \in (\overline{\lambda}(\mu_o)-\varepsilon\ ,\overline{\lambda}(\mu_o)+\varepsilon)$. That is, $\mu \in (\mu_o-\delta,\mu_o+\delta) \Rightarrow \overline{\lambda}(\mu) \in (\overline{\lambda}(\mu_o)-\varepsilon,\overline{\lambda}(\mu_o)+\varepsilon)$. Equivalently,

$$|\mu-\mu_o| < \delta \Rightarrow |\overline{\lambda}(\mu)-\overline{\lambda}(\mu_o)| < \varepsilon.$$

This means that $\overline{\lambda}$ is continuous in μ at $\mu_o \in (\mu^*,\mu^{(1)})$.

Continuity at μ^* is proved in a similar way. Thus $\overline{\lambda}$ is continuous in μ on $[\mu^*,\mu^{(1)})$. We conclude that $\overline{\lambda}$ is continuous in μ on $(0,\mu^{(1)})$, as stated in the lemma. □

CHAPTER V

THE DISTRIBUTION OF MAXIMAL QUEUE LENGTH IN THE M/G/1 QUEUE

We consider the M/G/1/n queue: customers arrive according to a Poisson process, with rate λ, at a system composed of a single server and n waiting positions. Service times are identically distributed, positive random variables, independent of the arrival process and each other, with distribution function H(x) and mean value $a = \int_0^\infty x dH(x)$. An arriving customer who finds the server busy and all n waiting positions occupied is cleared from the system; all others wait as long as necessary for service. The order of service is not specified.

This model, with explicit specification of the maximum allowable queue size, is important in applications because it allows one to examine the relationship between the number of waiting positions provided and the proportion of customers who will be denied service. The finiteness of the size of the waiting room makes the analysis of this model more difficult than that of its infinite-waiting room counterpart (see, for example, Cohen [1969], Cooper [1972] and Riordan [1962]). Of interest in this context are the mean duration of the busy period as a function of the number n of waiting positions, and the distribution of maximal queue length during a busy period in the corresponding system with an unlimited number of waiting positions. In this paper we show that these two quantities are intimately connected. In particular, we use this observation to augment

a theorem of Takács, namely Theorem 3 of Takács [1969], which we now state. In what follows, we adopt the notation and terminology of Takács [1969].

Let $\xi(t)$ be the queue size at time t, that is, the total number of customers in the system at time t. $\xi(0)$ is the initial queue size, that is, the number of customers already waiting for service at time t=0. Let θ_0 be the length of the initial busy period, and for $0 \leq i \leq k$ define

$$P(k|i) = P\{ \sup_{0 \leq t \leq \theta_0} \xi(t) \leq k \mid \xi(0)=i \} \tag{1}$$

as the probability that the maximal queue size during the initial busy period is $\leq k$ given that the initial queue size is i. Let π_j be the probability that exactly j customers arrive during a service time; then

$$\pi_j = \int_0^\infty e^{-\lambda x} \frac{(\lambda x)^j}{j!} dH(x) \qquad (j=0,1,2,\ldots), \tag{2}$$

with generating function $\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j$ given, for $|z| \leq 1$, by

$$\pi(z) = \psi(\lambda - \lambda z), \tag{3}$$

where $\psi(s)$ is the Laplace-Stieltjes transform of the service-time distribution function,

$$\psi(s) = \int_0^\infty e^{-sx} dH(x). \tag{4}$$

Then Takács's theorem (Theorem 3 of Takács [1969]) is as follows:

Theorem. For $0 \leq i \leq k$ we have

$$P(k|i) = \frac{Q_{k-i}}{Q_k} \tag{5}$$

where

$$Q(z) = \sum_{k=0}^{\infty} Q_k z^k = \frac{Q_0 \pi(z)}{\pi(z)-z} \tag{6}$$

for $|z| < \delta$ and δ is the smallest non-negative real root of

$$\pi(z) = z. \tag{7}$$

If $\lambda a \leq 1$, then $\delta = 1$ and if $\lambda a > 1$, then $\delta < 1$. Q_0 is an arbitrary non-null constant.

In his proof, Takács shows that

$$Q_k = \sum_{j=0}^{k} \pi_j Q_{k+1-j} \qquad (k=0,1,2,\ldots), \tag{8}$$

from which (6) follows immediately. (The proof given in Takács [1969] is more elementary than his earlier proofs--see the references in Takács [1969]). Cohen [1967], [1969] has also studied the distribution of maximal queue length; his results (see pp. 252, 571-2 of Cohen [1969] appear to be more complicated than those of Takács.) Note that the probability $P(k|i)$ is the same whether the waiting-room size n is finite or infinite, as long as $n+1 > k$. It is worth remarking here that, with $P(0|0) = 1$, (5) implies $P(k|i) = P(k|k)/P(k-i|k-i)$; thus Q_k may be given the interpretation $[P(k|k)]^{-1}$.

In this paper we show that

$$P(k|i) = \frac{b_{k-i}}{b_k}, \tag{9}$$

where b_n is the mean busy period in the M/G/1/n queue; that is, if we take $Q_0 = b_0 = a$, then $Q_n = b_n$.

1. The Mean Busy Period

We begin with the following recurrence on n for the M/G/1/n queue:

$$b_n = a + \sum_{j=1}^{n-1} \pi_j \sum_{k=n-j+1}^{n} b_k + \sum_{j=n}^{\infty} \pi_j \sum_{k=1}^{n} b_k \qquad (n=0,1,2,\ldots), \tag{10}$$

with the convention that any undefined sum is taken to equal zero. To prove (10) observe first that, clearly, $b_0 = a$. Now assume $n \geq 1$. Observe that the busy period is composed of the service time of the first customer plus some additional time if there are any new arrivals during the first service time. Suppose that exactly j ($1 \leq j \leq n-1$) arrivals occur during the first service time. Then, as the second service time begins, there will be j-1 customers waiting in the queue. Since the length of the busy period does not depend on the order of service of waiting customers, we can imagine that none of these j-1 waiting customers will enter service until any and all new customers are served who enter the waiting room after the start of the second service time. Thus the mean time until the next (if there is one) of the original j customers enters service is b_{n-j+1} (because only n-j+1 waiting positions were available to new arrivals during this time). Hence, using this queue discipline, the mean time required to serve all of the original j waiting customers is $b_{n-j+1}+\ldots+ b_n$; this explains the second term on the right side of (10). Finally, if $j \geq n$ customers arrive during the first service time, then the mean time until the completion of service of those n customers who enter the waiting room during the first service time is $b_1+\ldots+ b_n$.

Equation (10) can be written

$$b_n = a + \sum_{k=0}^{n-1} (1 - \sum_{j=0}^{k} \pi_j) b_{n-k} \qquad (n=0,1,\ldots). \tag{11}$$

If we subtract the (n-1)th equation from the nth in the above set, we get the following system, which appears most suitable for calculation of $b_1, b_2, \ldots$ by recurrence:

$$b_n = \begin{cases} \pi_0^{-1} a & (n=1) \\ \pi_0^{-1} \left[(1-\pi_1) b_{n-1} - \sum_{j=1}^{n-2} \pi_{n-j} b_j \right] & (n=2,3,\ldots). \end{cases} \tag{12}$$

Observe that Equation (12) can be written

$$b_k = \sum_{j=0}^{k} \pi_j b_{k+1-j} \qquad (k=0,1,\ldots;\ b_0=a). \tag{13}$$

Equation (13) is identical to Equation (8); therefore, Equation (9) is true as asserted, and further, if we take $Q_0 = a$ then $Q_n = b_n$ $(n=0,1,2,\ldots)$.

Similar results have been obtained by Tomko [1967], Cohen [1971] and Rosenlund [1973].

2. 'Direct' Derivation of Equation (9)

We have shown that the distribution of maximal queue length is a ratio of mean busy periods, but this equality appeared as a consequence of the fact that the quantities b_n and $Q_n (n=0,1,2,\ldots)$ satisfy the same recurrence (8). We now show that (9) can be obtained directly from

arguments that relate only to the $b_n (n=0,1,2,\ldots)$.

We begin by defining the i-busy period as the continuous busy time of a server that starts serving when i customers are in the system. (The 1-busy period is thus the ordinary busy period.) Let $B_k(i)$ be the duration of the i-busy period in the M/G/1/k queue $(1 \leq i \leq k+1)$, and define $B_\infty(i) = B(i)$, $B_k(1) = B_k$, $E[B_k(i)] = b_k(i)$, and $b_k(1) = b_k$. (Note that, in Takács's notation, $\theta_0 = B_\infty(i)$ when $\xi(0) = i$.)

Now, it is clear that

$$B_k(i) = B_{k+1-i} + B_k(i-1) \qquad (k=1,2,\ldots;\ i=2,\ldots,k+1), \tag{14}$$

from which it follows that

$$b_k(i) = b_{k+1-i} + b_{k+2-i} + \ldots + b_k \qquad (k=0,1,\ldots;\ i=1,\ldots,k+1). \tag{15}$$

It is also true that

$$b_k(i) = b_{k-1}(i) + [1 - P\{\sup_{0 \leq t \leq B(i)} \xi(t) \leq k \mid \xi(0) = i\}]b_k \quad (1 \leq i \leq k). \tag{16}$$

To prove (16) observe that during the i-busy period in the M/G/1/k queue we can imagine that the customer C, if any, whose arrival causes the waiting room to be fully occupied for the first time, will not enter service until there are no other waiting customers. Then the mean time from the start of the i-busy period until the system is cleared of everyone but C (if he exists) is $b_{k-1}(i)$. If in fact no such customer C arrives during $[0,B(i)]$, then the i-busy period ends; if C does arrive, which occurs with probability $1 - P\{\sup_{0 \leq t \leq B(i)} \xi(t) \leq k \mid \xi(0)=i\}$, then the additional mean time required to serve C and all his descendents is $b_k(1) = b_k$.

Equations (15) and (16) together imply (9). Thus, we have given a 'direct' proof of (9), as promised.

BIBLIOGRAPHY

Adiri, I. and Yechiali, U., "Optimal Priority-Purchasing and Pricing Decisions in Nonmonopoly and Monopoly Queues," Operations Research, Vol. 22 (1974), pp. 1051-1066.

Balachandran, K. R., "Purchasing Priorities in Queues," Management Science, Vol. 18 (1972), pp. 319-326.

Balachandran, K. R. and Lukens, J. C., "Stable Pricing Policies in Service Systems," forthcoming in Zeitschrift für Operations Research (1976).

Balachandran, K. R. and Schaefer, M. E., "Public and Private Optimization at a Service Facility with Approximate Information on Congestion," College of Industrial Management, Georgia Institute of Technology, 1975.

Balachandran, K. R. and Schaefer, M. E., "Regulation by Price of Arrivals to a Congested Facility," College of Industrial Management, Georgia Institute of Technology, 1976.

Cohen, J. W., "Distribution of the Maximum Number of Customers Present Simultaneously During a Busy Period for the Queueing Systems M/G/1 and G/M/1," Journal of Applied Probability, Vol. 4 (1967), pp. 162-179.

Cohen, J. W., The Single Server Queue, North-Holland, Amsterdam, 1969.

Cohen, J. W., "On the Busy Periods for the M/G/1 Queue with Finite and with Infinite Waiting Room," Journal of Applied Probability, Vol. 8 (1971), pp. 821-827.

Cooper, R. B., Introduction to Queueing Theory, Macmillan, New York, 1972.

Cooper, R. B. and Tilt, B., "On the Relationship Between the Distribution of Maximal Queue Length in the M/G/1 Queue and the Mean Busy Period in the M/G/1/n Queue," Journal of Applied Probability, Vol. 13 (1976), pp. 195-199.

Crabill, T. B., Gross, D., and Magazine, M. J., "A Survey of Research on Optimal Design and Control of Queues," Serial T-280, School of Engineering and Applied Science, George Washington University, 1973.

Henderson, J. M. and Quandt, R. E., Microeconomic Theory: A Mathematical Approach (2nd ed.), McGraw-Hill, New York, 1971.

Intriligator, M. D., Mathematical Optimization and Economic Theory, Prentice-Hall, New Jersey, 1971.

Kleinrock, L., "Optimum Bribing for Queue Position," Operations Research, Vol. 15 (1967), pp. 304-318.

Kleinrock, L., Queueing Systems, Vol. I: Theory, John Wiley & Sons, New York, 1975.

Knudsen, N. C., "Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure," Econometrica, Vol. 40 (1972), pp. 515-528.

Littlechild, S. C., "Optimal Arrival Rate in a Simple Queueing System," International Journal of Production Research, Vol. 12 (1974), pp. 391-397.

Luce, R. D. and Raiffa, H., Games and Decisions, John Wiley & Sons, New York, 1957.

Naor, P., "The Regulation of Queue Size by Levying Tolls," Econometrica, Vol. 37 (1969), pp. 15-24.

Riordan, J., Stochastic Service Systems, John Wiley & Sons, New York, 1962.

Rosenlund, S. I., "On the Length and Number of Served Customers of the Busy Period of a Generalised M/G/1 Queue with Finite Waiting Room," Advances in Applied Probability, Vol. 5 (1973), pp. 379-389.

Stidham, S. and Prabhu, N. U., "Optimal Control of Queueing Systems," Mathematical Methods in Queueing Theory, Springer-Verlag, Berlin, 1974, pp. 263-294.

Takács, L., "On Inverse Queuing Processes," Zastosowania Matematyki, Vol. 10 (1969), pp. 213-224.

Tomko, J., "A Limit Theorem for a Queue When the Input Rate Increases Indefinitely," (In Russian.), Studia Scientiarum Mathematicarum Hungarica, Vol. 2 (1967), pp. 447-454.

Yechiali, U., "Customers' Optimal Joining Rules for the GI/M/s Queue," Management Science, Vol. 18 (1972), pp. 434-443.

VITA

Borge Tilt was born on 21 September, 1932 in Copenhagen, Denmark. He attended Sct. Jorgens Gymnasium, graduating in June 1951. Thereupon he entered the School of Economics, University of Copenhagen, receiving the degree cand. polit., an M.A. in Economics and Statistics, in June 1959.

From September 1960 to December 1963 he was a statistician for Danfoss Ltd., Nordborg, Denmark working in the areas of experimental design and analysis, quality control, and operations research. From 1964 through 1965 he was a consultant in quality control.

From June 1966 to June 1971 he was employed with the Lockheed-Georgia Company, Marietta, Georgia. His work experience in the Scientific Computing Division included computer operations scheduling, simulation studies, and forecasting of air cargo load.

He went back to school in September 1971, earning the degree of Doctor of Philosophy in the College of Industrial Management, Georgia Institute of Technology, in August 1976. He will return to Denmark to work as a management consultant.

He has published papers in Danish in queueing theory and statistics, and has coauthored a paper for the Journal of Applied Probability.

In November 1973 he was married to Seiko Koyama of Sendai City, Japan. They have a daughter, Yumi.